DOCUMENT RESUME

ED 233 326 CS 007 270

AUTHOR Singer, Harry; And Others
TITLE Testing Achievement in Basic Skills. Volume II: Description, Analysis, and Recommendations. Final Report.
INSTITUTION California Univ., Riverside. School of Education.
PUB DATE Jun 83
NOTE 167p.; Project funded by the California Policy Seminar. For related document see CS 007 269.
PUB TYPE Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC07 Plus Postage.
DESCRIPTORS Academic Standards; Achievement Tests; *Basic Skills; Criterion Referenced Tests; Educational Research; Educational Strategies; Educational Testing; Elementary Secondary Education; Norm Referenced Tests; Program Evaluation; Reading Instruction; Standardized Tests; *State Programs; *Test Construction; *Testing; *Testing Problems; Testing Programs; *Test Reviews

ABSTRACT

Intended both to evaluate the California testing program's success in assessing student achievement in the basic skills and to suggest possible improvements in the program, this report includes an executive summary evaluating current testing procedures, describing experimental studies on new ways of testing and reporting testing information, and making recommendations for improving the current program; an overview of the testing program, the specific questions posed for this report, and a list of criteria and assumptions for evaluating tests; and a review of tests of basic skills at the state, school district, and classroom levels and applies the report's evaluation procedures are applied to these tests. Descriptions of norm and criterion referenced tests are included in the appendix. (MM)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
X This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

ED233326

Rev. 6/83

TESTING ACHIEVEMENT IN BASIC SKILLS

Volume II: Description, Analysis, and Recommendations

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
Harry Singer
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Harry Singer
Riverside

Robert Ruddell
Berkeley

John McNeil and Merl C. Wittrock
Los Angeles

University of California

This report was prepared for the California Policy Seminar on a grant awarded for this purpose, March 1981, and administered by the Institute of Governmental Services (John Cummins, Director), University of California, Berkeley.

CS 007270

Table of Contents

	Page
Foreword	1
Introduction	5
Executive Summary	7
Testing Achievement in Basic Skills	7
Need for This Study	8
Current Testing Programs	10
Research on Tests and Testing	12
State Testing	12
School District Testing	13
Classroom Testing	15
Recommendations	17
Statewide Testing Program	17
District Level Testing	24
Classroom Level Testing	29
Footnote	31
Overview of Testing Program	32
Constrasting Uses of Educational Testing	33
Assessment Assumptions	36
State Testing: The California Assessment Program	46
CAP Uses	56
CAP Limitations	57
Conclusion	69

	Page
School District Testing	74
Classroom Testing	117
Uses and Limitations of Tests and Test Results	119
Ideal Types and Uses of Tests	124
References	133
Appendix	143
Norm-Referenced and Criterion-Reference Tests	144
Addendum: Changes in Pupil Proficiency Laws, Failure Rates, and Estimated Costs of Mandated Remedial Proficiency Classes	156

Tables

		Page
1	Tests in the California Assessment Program	49
2	Summary of Background Factors Used for Computing Comparison Score Bands for 1974-75 and 1975-76	54
3	Standardized Norm-Referenced Achievement Tests for Basic Skills Whose Norms Are Used by the California Assessment Program Tests	85
4	Paragraph with Accompanying Comprehension Question from the Metropolitan Reading Instructional Tests, Elementary, Form J	90
5	Descriptions of Two Proficiency Tests	105
6	Condensed Norms for Grade 5 (Fall)	147

Figures

1	Example of National Assessment of Education's Report of Test Results	43
2	Comparison of Student and School Percentile Ranks	63
3	Repeated and Grade Level Testing	89
4	Measured Range of Reading Comprehension Ages in Grades 2 through 8	91
5	Diagnostic Information on an Hypothetical Student's Progress in Learning How to Read and in Use of Reading to Gain Information	93
6	Current Testing and Report of Numerical Information, Content-Free Contrasted with a Combination of Self- and Peer Group Comparison Testing with Content-Bound Information	97

Foreword

Although educators throughout history have used tests of one form or another, only in the last 100 years have standardized, objective tests been developed and employed systematically on a large scale in education. This phase of testing began in reaction to the excesses of the spoils system following the Civil War. Congress voted to establish a civil service which began to devise tests for the objective selection of candidates without regard to race, religion, or social origin. In 1897, a spelling test became the first standardized achievement test. Over the next 20 years enough achievement tests were devised to provide instruments for the scientific movement in education that promised to make schools more efficient by grouping students for more appropriate instruction. Another type of test, one for assessing intelligence, developed by Binet and Simon in 1905-1911 was imported, improved by Terman at Stanford in 1916, named the Stanford-Binet, and used to assess students' capabilities for school achievement.

New York has used its own Regents Examination for over 100 years to assess achievement at the high school level.[a] But California did not begin to use state mandated tests until 1961. Over the next 20 years it employed tests to assess achievement in basic skills, monitor local school districts' curricula, certify competency of students and teachers, and use the information to determine allocation of resources.

Tests have added to the State's control over education. The control already included establishment of local school boards to administer the State's Education Code; textbook selection; credential

requirements for teachers, administrators and specialists; and funds for categorical programs. Most recently, after the passage of Proposition 13, the state began to provide bail-out money for local school districts who could no longer depend upon property taxes as their main source of support. Thus the state legislature has assumed greater control over schools at the cost of local control.

However, this report focuses on only one aspect of legislative control over education: testing achievement in basic skills. But the scope of our inquiry is broader. We include information on testing required by the federal government, the school district, and classroom teachers. We also touch on aptitude and achievement tests required for admission to the state's colleges and universities. We do not critically evaluate the college entrance tests, but we point out that a gap exists in information gathered at the high school level for improving the testing program because the state only monitors achievement in basic skills and requires minimal competency testing, but does not assess achievement in college preparatory programs.

Concern over the testing program led the California Policy Seminar to commission a study on "Testing Achievement in Basic Skills." The study consists of two parts: (1) a description and analysis of the current testing program and (2) research on several aspects of achievement testing. We reported the results of the study in two volumes. Each volume contains the same Foreword and Executive Summary. This volume emphasizes description and analyses of the current testing program. The other volume contains the research reports. All but one of the research reports have been submitted for publication in professional journals.

In working on this study we have received help from many people. Members of the California Assessment Program, Drs. Alex Law, Dale Carlson, Beth Breneman, and Mark Fetler, were most generous in explaining the CAP program and providing us with tests, materials, reports, and even statistical analyses of CAP data. If we err in this report on any facts pertaining to the CAP program it is not due to any unwillingness on the part of the CAP staff to provide us with the necessary information. We also want to thank the many legislators, their staffs, and the public school personnel and students who responded to our questionnaires, gave us interviews, and participated in our experimental activities for determining test preferences and attitudes towards testing and test reporting. We also received help from the San Diego Unified School District. Dr. Bob Bennett, director of curriculum, provided data which showed that San Diego had expanded its time on task in basic skills and thereby had narrowed its elementary curriculum over the last 5 to 10 years. Dr. Steve Isaacs, San Diego's Director of Research and Evaluation, gave a lengthy and detailed response to our request for a school district's view on testing; much of his response is reported in our Final Report, Volume II, Section III. Mr. Walter Lansu, now retired from the Los Angeles Unified School, and his staff informed us of how the minimal competency testing program would operate in their district. Dr. Robert Hocker, now acting Superintendent of Alvord Unified School District, made arrangements for testing and communicated considerable information about the CAP program and school district testing. Dr. Ross Green of McGraw Hill/California Test Bureau provided us with his company's tests

and other information on tests and testing. Dr. Starrett Dalton, Director of Research and Evaluation for the Riverside Unified School District, and Mr. Ray Berry, formerly superintendent of that district and now lecturer in Education at the University of California, Riverside, informed us of testing practices in Riverside. We are also grateful to the colleagues, research assistants, and secretaries who worked on aspects of this report. They are Dean Irving H. Balow and Dr. Douglas Mitchell, University of California, Riverside; Dr. Jean Dreher, University of Maryland, College Park; Dr. Charles Kinzer, now at Vanderbilt University; Richard Speaker and Marilyn Draheim, graduate students at the University of California, Berkeley; and Lory Furth and Helen Schultze, graduate students at the University of California, Los Angeles. Thanks also goes to Betty Medved who typed the original manuscript and to Linda Faulkner who willingly typed each of our revisions until we finally reached consensus on all aspects of the report. We also appreciate the careful, considerate, and cordial administration of our budget by Rosie Russell, Management Services Officer. Finally we express our gratitude to the California Policy Seminar for awarding us this project and Mr. John Cummins and his staff at the Institute of Governmental Services that oversees the project.

We dedicate our report to all the students in California's schools; we hope that the report will help make their education better through improvements in tests and testing practices at the state, school district, and schoolroom levels.

HS, RR, JM, and MW

August, 1982

Introduction

Evaluating educational achievement is necessary because the knowledge generated from testing students' achievement helps all those concerned with public education make informed and intelligent decisions about education. For example, it provides information to help answer such questions as how should the state best allocate its limited resources so as to improve achievement in basic skills in particular and education in general? Did the state's increase in funds to a particular school district improve its students' performance? Is one educational program more effective than another in achieving some educational objective? What effect did a change in state law, for example, from state mandated textbooks to local adoption of textbooks from a state list of textbooks in compliance with educational and legal requirements, have upon the quality and level of school achievement? These and other questions require the state and schools to collect and analyze test data in order to even approach an answer to the questions.

However, our purpose is not to question the necessity for testing and evaluating. What we inquire about in this report is whether the program for assessing achievement in basic skills is attaining its intended purpose(s) and whether any aspect of it can be improved. As the report will show, we think that the tests, the ways in which the state, schools, and teachers administer and use them, and particularly how test data are reported to the public can and should be changed in ways that will be more beneficial to all those concerned with public education.

The report is divided into five sections. Section I contains an Executive Summary. Section II contains an overview of the testing program, the specific questions posed for this report, and a list of criteria and assumptions for evaluating tests. In Sections III to V we review tests of basic skills at the state, school district, and classroom levels and apply our evaluation criteria to them.

D50

Section I

Executive Summary

Testing Achievement in Basic Skills[b]

The purpose of this report is to evaluate the current testing program, conduct experimental studies on new ways of testing and reporting test information, and make recommendations on how the testing program can be improved in California. As the report shows, we think that the tests, the ways in which the state administers and uses them, the confidence placed in the validity of test information, and how test data are reported should be changed in ways that will make achievement testing in basic skills more beneficial to all those concerned with public education. The evidence for our conclusions and the basis for our recommendations stem from descriptions and analyses of tests currently used in California schools, reports on tests and test results for California students, and seven studies that we conducted for this report.

This summary consists of four parts. The first is an explanation of the need for doing this study, the second is a description of tests in current use, the third contains summaries of our research results, and the fourth is a list of our recommendations.

Need for This Study

Tests not only assess pupils' progress and performance. they also serve for selection and placement of students in programs, and influence students' self-perceptions. Moreover, they reflect upon the competence and efforts of professional educators. Tests operationally define the goals of education, tend to control and regulate what is taught and emphasized in the curriculum, and provide information that influences public policy decisions on education. They also inform the public on the efficiency and effectiveness of the schools in attaining educational goals. Hence tests serve schools and society in many ways.

Testing achievement in basic skills is also an expensive enterprise. California's school districts spend over 40 million dollars each year on assessment. This dollar amount does not include the prorated costs of teacher and student time nor any other indirect costs.

Teachers at best are ambivalent about the testing program. They perceive that the tests use up expensive instructional time. They know that the state-wide testing program cannot provide individual test scores to teachers because each student takes only a sample of test items, not a complete test. Hence the state can only return information on grade levels, schools, districts, and state trends on basic skills. This information cannot be related to individual students. Although the information may be useful to the state, the district, and school principals, it has little meaning for classroom teachers.

Information on the district's standardized tests, which are usually scored at some central location, takes about 3 months to reach teachers. After this time interval, teachers do not have much interest in the test data, particularly in primary grades where student development in achievement is relatively rapid. However, standardized test information from commercially published tests, in contrast to statewide data, informs school boards, parents, principals, and pupils on achievement in basic skills in comparison with national norms. The information that teachers find most instructionally relevant is obtained from the results of their self-constructed or criterion-referenced tests. Students know these tests are related to their grades and the results will be returned to them; therefore, they are likely to be motivated to do their best.

A criterion-referenced test has items or questions related to a particular instructional objective. Ideally students should know the answer to few items before instruction and most items after instruction. The teacher sets a criterion of the percent of items students must get correct to indicate they have learned the objective. Thus, a criterion-referenced test is designed to be instructionally sensitive.

In comparison, a standardized, norm-referenced test also consists of items or questions related to a particular objective(s), but the test-maker, usually a commercial publisher, then gives the test to a sample of students and does an item analysis of the results. Those items are eliminated that all students fail or pass. Standardization consists of retaining only those items that fit certain statistical

criteria. When the test is given to another sample of students, the resulting scores form the comparison norms for the test.

Because of the multitude of tests California students now take, the significant role they play in school and society, the large sums of money expended on them, and the ambivalent attitude teachers have toward tests, they need to be scrutinized objectively and critically.

Current Testing Programs

California students take state-mandated, school district, and teacher-made tests from grades 1-12. The state-mandated test, constructed and administered by the State Department of Education's California Assessment Program (CAP), is called the Survey of Basic Skills (SBS).[3,4,5] It is given annually to each public school student enrolled in grades 1, 3, 6 and 12. School districts must also test students on nationally standardized tests constructed by various publishers. The most widely used standardized tests in California are the California Test Bureau/McGraw Hill's Comprehensive Test of Basic Skills, the Psychological Corporation's Stanford Achievement Test, and the Science Research Associate's Iowa Tests of Educational Development. School districts must administer these kinds of standardized tests to satisfy requirements for students enrolled in federally-funded Title I programs. Some districts also administer standardized tests annually to all its students in each grade. Other districts such as the Los Angeles Unified School District administer

its own test of basic skills to evaluate student progress. Students in all schools also take teacher-made tests which assess achievement of specific instructional objectives. Those high school students who want to apply to the California State University and College system or the University of California must take the Educational Test Service's Scholastic Aptitude Tests. For the University of California, they must also take the College Entrance Examination Board's Achievement Test(s).[1]

Another state-mandated test that all California students must pass to receive a high school diploma is called the proficiency or minimal competency test.[13] It assesses achievement in reading, writing, and arithmetic. According to state law, school districts must select or construct their own proficiency tests, decide the score necessary for passing the test, and administer them once each year in grades 4-6, 7-9, and 10 and 11. Students who pass the tests in lower grades do not need to take the tests again.

The kinds of questions we sought to answer in this report about the testing program are: What use is made of the test results? Do they help legislators decide how to allocate resources? Are they useful to administrators and teachers? Are there other ways of teaching, testing, and reporting test results that would better serve the purposes of education in California? To answer these questions we not only read and evaluated published reports on the testing program, we also conducted seven research studies.

Research on Tests and Testing

State Testing

Interviews of a representative sample of educators, including administrators and teachers throughout the state, revealed that as the interviews progressed from the district to the classroom level, the less was known about the CAP test, how to interpret its results, and the usefulness of its information. Administrators tended to report the test data were somewhat useful, but wanted fewer skills tested; teachers indicated the results were only of marginal use or not useful at all.[7,8,14,15]

Legislators and members of their staff also had an ambivalent attitude toward the CAP program. They reported that while educators are responsible for achievement in basic skills, legislators need achievement information for making decisions on allocation of funds to schools. Most legislators interviewed, however, have little confidence in the objectivity of the state test results. They recognize that the State Department of Education which is responsible for administering and providing leadership for the state's educational program has an elected superintendent whose Department of Evaluation and Assessment is in charge of all aspects of the CAP test from construction of the test to interpretation and release of the results to the public and the press. In effect, the Superintendent is required to evaluate his own program. Legislators realize that this requirement may lead to a conflict of interest. Because of this fact alone, some legislators do not have confidence in the test results and

their interpretation. Several legislators recommended that a contract should be made with an independent group to administer and interpret the assessment program.[14]

School District Testing

Under the Educational Code, school districts must select or construct their own proficiency test. Some 70 percent of the school districts in California elected to construct their own test. Acting under guidelines from the State Department of Education, school districts constructed tests and decided upon a passing score that resulted in some 30 percent or 100,000 10th graders failing one or more proficiency tests in 1981, the first year the test had to be administered under state law. School districts estimated that some 5 percent or 30,000 students will still fail the test when they reach the 12th grade; consequently they will not receive a high school diploma.[13] They will therefore not be eligible for any job that requires one. Since the law requires schools to provide instructional programs for students who fail the test and continue to provide them until the students pass, schools had to reallocate resources to meet this requirement. An estimated $12,000,000 in school funds had to be shifted from the 70 percent of the school population who passed the tests to the 30 percent who failed one or more of them.[16] Since many of the tests involve application of basic skills to life-tasks, such as filling out a rent receipt, reading an advertisement, or computing a grocery bill, that are not usually taught as such in school,

questions have been raised about the legality of the tests.[10] Already some parents have taken school districts to court over due process issues raised by the tests.

School districts also administer standardized tests of basic skills. These tests are well-constructed, provide reliability and validity data, and have national norms. However, these norm-referenced scores are difficult for teachers to use. We conducted an experiment which demonstrated that teachers assign different letter grades to identical student scores that were first reported in percentiles and then in percentages. Teachers seem to use a 100 point scale for evaluating test results as though both types of information indicate percentage of items correct. In other words, they react to the numbers alone, not to what they mean. This experiment indicates that teachers need student scores reported in a familiar and understandable format which will be useful for their diagnostic decisions and instructional plans.[22]

School districts do not report test results in a meaningful way. They usually only indicate percentile rankings from year to year. Lines connecting these rankings are usually flat unless a student has improved more rapidly than the norm group, then the line slants upward. A student may improve but at a slower rate than the norm group, then the line slants downward. These types of reports are susceptible to misinterpretation. In all three cases, the student may have improved, but the public would interpret only the upward slanting line as an improvement. Moreover, districts and schools report scores in numerical form only. Because such reports are content-free, they do not inform the public on what students know and can do. When

percentile rankings are given on the same test at the beginning and end of the year and compared with the same norm group, then a line connecting the resulting percentile rankings is likely to show an upward slant, particularly in the primary grades where basic skills are being acquired at a rapid rate.

These rankings should also be shown with content-bound information, for example, the most difficult paragraph a student could read successfully each time the student was tested. If such information were reported, the public, press, parents, and pupils would perceive primary grade students were developing rapidly in acquisition of basic skills. If such self-comparison data were also accompanied with the usual peer-comparison data and with content-bound information, educators would find the information is diagnostically useful and would change their attitudes to testing from negative to positive. In short, educators are not opposed to testing as such, but to the way that tests are currently administered and reported.[17,18]

Classroom Testing

We found from questionnaire data for third and sixth grade that a majority of teachers indicated tests are important and use test results for making relevant decisions. However, the teachers reported that individual test data are most useful for making classroom decisions.[22] Teachers can only obtain this information from district administered standardized tests and teacher-made tests, not from the state administered tests.

Teachers prefer test data that are related to what their instructional materials emphasize and what standardized tests assess. They also want information on the newer processing strategies for attaining reading comprehension. In short, teachers prefer coherence between what they teach and what they test.[22]

Teachers also favor criterion-referenced to norm-referenced data. They tend to perceive students more favorably when their scores are reported in a criterion-referenced format, in percentage of correct items.[22]

To determine the kind of diagnostic information and the type of reporting format teachers prefer we gave teachers and administrators four types of test information on the same second grade student. They most preferred the diagnostic report that differentiated rate of reading acquisition from peer group comparison and provided content-bound information, that is, the most difficult paragraphs the student could read successfully each time the student was tested. The educators least preferred the report which listed peer group comparisons only, the type of report that is usually given for standardized testing![17] An attitude scale indicated educators had a favorable attitude toward the diagnostic report, but a negative attitude toward the peer group comparison only type of report.[18]

Because instruction and testing are interrelated, we conducted an experiment to determine what effect teaching one of the newer comprehension processing strategies would have on the state's test of comprehension. We taught elementary students how to summarize what they read. They improved significantly over another group that was asked to summarize what they had read before they were tested, and both groups achieved more on the CAP comprehension items than a group which only read the passage and then answered the questions. In short, more appropriate comprehension instruction resulted in significant improvement on the comprehension test.[11]

Our analyses of current test practices, test reports, and our own experimentation with testing have led us to formulate recommendations for the improvement of achievement testing of basic skills at the state, school district, and schoolroom levels. We follow each recommendation with a brief rationale.

Recommendations

Statewide Testing Programs

Rec. 1. Retain, but modify CAP's Survey of Basic Skills Test for

monitoring state, district, and school performance but include newer types of process oriented skills. Administer CAP to only a small sample of representative schools within school districts each year. Change the title of the CAP report from "Student Achievement" to "School Achievement." Authorize an independent agency, not the State Department of Education, to direct, administer, and interpret the statewide assessment program.

Each administrative level must collect appropriate information for use in making intelligent decisions. The State Department of Education must gather data on how California as a state is performing relative to other states in achievement of basic skills (Education Codes 60600, 60663 and 60640). Although the CAP test meets the criteria for a well-constructed test, it has its strengths[9] and limitations.[2] Its major strength is to provide the state, school districts, and schools with comparative data on achievement in basic skills. Its major limitation is that it does not provide individual scores; even if it did administer individual tests, the Educational Code (60607, 60641) forbids annual reporting of state-determined individual scores or relative positions of individual students to district governing boards; however, the State Board of Education may determine which, if any, of the achievement test results in grades 3, 6, and 12 may be recorded on students' cumulative records (Education Code

60612, 60641). However, the Education Code (60605) permits school districts to administer other tests, including tests that yield individual scores.

The CAP tests also have other limitations. We have already discussed its need to incorporate a more effective way of measuring processing skills. The CAP tests have been recently broadened at the sixth grade to include reports not only on basic skills but also on performance in content areas of science, social studies, and literature.[5] This broadening is likely to counteract a tendency for elementary schools to narrow their curricula to emphasize only basic skills. This emphasis may have been counterproductive to achievement in basic skills, particularly to reading comprehension.[12] A similar broadening of the CAP tests should be done when the 12th grade tests are revised.

The CAP test also lacks national norms. Unfortunately, there are no test norms on a state-by-state basis. The only way such norms could be constructed is to administer the CAP tests to a representative sample of schools in each state. We do not foresee such norms forthcoming. To approximate a comparison with other states, California equates its test to nationally standardized test results, but caution must be observed in interpreting these results because the theoretical narrowness of state-by-state norms would magnify greatly any state's departure of test results from the average on the hypothetical state-by-state norms.[19]

To collect information on state trends, it is not necessary to test every student in grades 1, 3, 6, and 12. The state could obtain the same result from a sample of schools in each district.[21] This sampling approach would be more economical than the current practice. Each school would then not be tested on the state test, but almost every school already tends to get similar data from standardized tests administered by local school districts.

The state could continue to collect and equate district standardized test results with the state test. The state currently uses such standardized test data to equate the CAP test with national norms. Most school districts are now using standardized tests to satisfy federal requirements for Title 1 programs.[19]

Our recommendations should be viewed in an historical context. In 1961 the legislature passed a law requiring the State Department of Education to provide data on achievement in California schools. The state implemented the law by collecting school districts' standardized test results. But difficulties in equating results from diverse tests led the state in 1962 to adopt a single standardized test for its statewide testing program. After some seven years of successive commercially-published test adoptions and use, the legislature responded to district objections over the validity of the tests by passage of the California School Testing Act of 1969 and by further legislation in 1972 that, among other changes, allowed the State Board of Education to construct its own test. The scores of the testing program were to be "limited primarily to the task of furnishing information useful to state policymakers and decision makers;"

it was assumed that "assessment information needed at the classroom level could best be collected by local school personnel. In spite of this assumption, the [CAP] program was designed to report as much information as possible to local personnel" (p. 6).[4] Our recommendation reinforces the legislature's limitation on the scope of the CAP program and advocates that the legislature's assumption be fulfilled. The CAP program would then only serve its intended purpose and would avoid problems that arise when it attempts to make its results useful to local personnel.

Although the state solved some problems in making the 1972 changes, such as increased test breadth and relevancy and reduced testing time, the change created other problems, such as having a test that does not provide individual scores and overlaps with school districts' standardized test results. Also, since the state continues to interpret its tests through the use of commercially published standardized test norms, it has not solved one of its original complaints that commercial tests have out-of-date norms. Our recommendation reduces the overlap, makes the state testing program more cost-effective, and solves the initial problem of how to compare results from diverse standardized tests. The state can now equate the standardized tests used in local school districts to the common CAP test.

The CAP tests should also be modified in another important way: they should be referred to as "School Achievement," not as "Student Achievement" data. Since no student takes a complete test and since no score can be reported on individual students, the term "student achievement" is misleading.[19]

More important is the lack of confidence of legislators in the CAP test results.[14] To spend millions of dollars on a testing program and not to have confidence in the results and their interpretation because of a possible inherent conflict of interest is cause enough for changing the CAP program. We therefore recommend that legislators modify the Educational Code in such a way that an independent agency can be contracted to administer the CAP program and satisfy the objectives of providing information to the legislature that is necessary for making decisions on allocation of resources, evaluation of programs, and identification of unusually successful practices and factors which appear to be responsible for attainment of high achievement (Education Code 60601).

Rec. 2. Devote resources of the State Department of Education to construction or selection of and experimentation with new tests and related materials and instructional procedures that may be useful for improving achievement in basic skills.

The State Department of Education has the resources and the capability to serve this function, which no one else is doing. For example, knowledge and procedures for assessing process-oriented skills, such as summarization have recently become available and have improved students' reading comprehension.[11] Assessment of this kind of skill will provide teachers with needed information on whether students are acquiring necessary ways of attaining achievement in basic skills.

Test publishers already know that process-oriented skills can be taught, learned, and assessed. But they are reluctant to incorporate the new skills in their test battery lest schools find the change too great and not purchase the test battery. Consequently, they only tend to assess current curricular and instructional objectives, as manifest in materials. Thus, they are assured their tests will have curricular and instructional validity. Teachers tend to teach what is in their instructional materials and want tests that assess what they have taught. To break this circular process, the State Department of Education will have to experiment with new materials that develop the new processes involved in attainment of achievement and use new tests to assess whether students are attaining the process. Apparently both tests and instructional materials have to change simultaneously for teachers to change what they teach and what they test or want to have tested.[22] Indeed, learning, instruction, and testing should be coherent.[20]

When the new tests are shown to be valid, they can be incorporated into the California Assessment Program's test battery. Thus, through curriculum innovation and assessment, the State Department of Education will be contributing to its leadership role in education.

Rec. 3. <u>The state legislature should gather information on the level and quality of performance of those students enrolled in college preparatory courses required for admission to the University of California, California State University and Colleges, and to other public and private universities.</u>

While analyzing the existing testing program, we became aware that information on school achievement that the state now obtains is inadequate for making decisions about allocation of resources at the high school level because the state does not monitor its mandated programs for college preparation. Apparently college preparatory standards in California are lower than the national average. For example, the College Entrance Examination Board reported that California students take fewer academic courses for less time and get higher grades than students in general throughout the nation.[1] However, if the legislature obtained achievement data on the college preparatory program by administering its own tests or relying on achievement test results from the College Entrance Examination Board tests, it could use such information for determining allocation of resources. It would then have information on all three tiers of high school students: those in the college preparatory program, those who still need to pass a minimum competency test, and those students in-between who represent the average performance in basic skills as assessed by CAP tests.

District Level Testing

Rec. 4. School districts should not only assess achievement in basic skills, but also in the content areas of science, social studies, and literature.

If school districts only assess achievement in the basic skills, teachers are likely to focus on development of these skills and thus narrow the curriculum. The result is that students would not attain as much general knowledge. This knowledge is necessary for reading comprehension. Most printed materials require students to draw upon general knowledge for making inferences as they read and for interpreting printed materials. The consequences of a narrow curriculum are more likely to be observed in the higher grades. The remedy is simple: assess students in both basic skills and in the content areas of science, social studies, and literature.

If standardized tests are used for this assessment, the results will also provide boards, parents, professionals, pupils, and the public with information on the relative standing of students, classes, schools, and districts in relation to national norms. We know that teachers need this information to help them make informed and objective decisions in selection and placement of students.[22] Fortunately, most districts are already providing this type of information.

Rec. 5. Districts should retest students in the primary grades on the same tests and use the same norms to interpret the scores.

Self-comparison data that results from retesting is likely to demonstrate that primary grade students generally make rapid progress

in acquisition of basic skills. This kind of information will inform legislators and the public on the degree of success that schools are experiencing in developing students' acquisition of basic skills.[17]

Rec. 6. Report all test results with sample tasks.

The sample tasks will enable the test consumer to see what students know and can do on each type of test at each level of testing.[17,18]

Rec. 7. The legislature should allocate funds for high schools to carry out the mandated remedial program for minimal competency attainment.

Given the authority, but with guidance from the State Department of Education, to construct or select their own minimal competency tests, school districts tended to set the standard at the ninth grade level with 70 percent as the passing score. Some 30 percent of the ninth grade students who fail the tests have to take one or more remedial classes at high school. School districts have had to reallocate their own resources for the mandated program since the state provides no additional funds for it.

However, some school districts set the standard for their minimal competency test at the sixth grade level where they were prior to minimal competency testing. They established a passing score at such

a level that at least 70 percent of the students in the sixth grade could pass the tests. Almost all students could pass these tests prior to high school. Consequently, their high schools had very few remedial courses in basic skills and almost all the students who completed 12th grade passed the test and received a high school diploma. Thus, these high schools did not have to reallocate much resources and do not need to seek much, if any, reimbursement funds.

If school districts want to adopt a sixth instead of a ninth grade standard, they should know that they already have the authority to make this change in their minimal competency tests.[16] These districts should also inform their boards that minimal competency requirements should not be confused with graduation requirements.

Rec. 8. The legislature should require each school district to adopt procedures for improving its administrators' and teachers' ability to (1) construct, administer, and interpret criterion-referenced and norm-referenced tests and (2) use test results for making decisions on curriculum and instruction. The State Department of Education should be authorized to provide school districts with guidelines on how to effect this improvement. These guidelines should be updated as new information on tests and testing becomes available. Each school district should then report to the State Department of Education how it is implementing this recommendation.

Our research has indicated that school district administrators and teachers need to have more knowledge on all aspects of testing[8,14,15] With more adequate knowledge, they will be better qualified to adopt or construct tests, administer them to students, interpret the results, and apply the knowledge gleaned from the tests to decisions-making processes in curriculum and instruction. Adequate knowledge of tests and testing should also enable them to beware of the pitfalls of assessment, particularly on test bias, in relation to minority groups.

We think that each school district should have the responsibility for deciding how it will implement this recommendation under guidelines from the State Department of Education. In providing guidelines the State Department should state the various ways in which this recommendation can be implemented and let each school district select an option or devise an alternative that will produce the same result. School districts should be encouraged to work with personnel in local colleges and universities in designing this plan. Each school district should then inform the State Department of Education of its decision and describe how it is implementing this recommendation.

School districts must also keep up with change in tests and testing as they occur. For example, new tests are likely to become available on assessment of knowledge a student has acquired, use of this knowledge, and diagnosis of what has been learned and what still needs to be improved in subject matter learning.[6] By incorporating

such information into revisions of its guidelines every few years, the State Department of Education will be helping local school personnel keep up-to-date on achievement testing in basic skills.

Classroom Level Testing

Rec. 9. Provide teachers with test data on individual students, both criterion and norm-referenced. These data should be diagnostic and reported in a meaningful way so that they can be used for instructional decisions at the classroom level.[17,18,22]

Teachers need criterion-referenced data for making instructional decisions. They also should have norm-referenced data for selection and placement of students.

If students, particularly in the primary grades, are retested on the same tests of basic skills after a year of instruction, as well as on a norm-referenced test standardized for the grade level, then diagnostic information would be available on each student's progress in acquisition of the skill and relative comparisons with a peer group.[17]

To make reports meaningful, they should contain information on percent of items correct and percentile rank. These two types of information are complementary. Reports for each student should also contain examples of the most difficult task the student successfully completed on each basic skill.[17,18]

The tests, both criterion and norm-referenced, should assess not only what students know but also how they acquire this knowledge. Hence, tests need to contain processing-oriented as well as knowledge scores.[11,22]

Elsewhere we have recommended that the resources of the State Department of Education should be devoted to construction of and experimentation with new types of tests and with preparation, selection, and try-out of new materials and instructional procedures to go with experimentation in new types of tests. To effect change in teaching of basic skills, it is necessary to modify both tests and materials simultaneously.

However, we would caution schools not to proliferate the number of criterion-referenced tests administered to students. Such proliferation could result from defining instructional domains narrowly and assessing each domain. For example, defining a domain of types of commas in punctuation, instead of a more general domain of punctuation. Only those domains should be assessed in which teachers need and will use information for making instructional decisions. All those involved with authority over testing students, from the state to the classroom level, need to exercise wisdom and judgment in exercising this authority.

With our recommendations in force, we believe that the state will have more adequate, more dependable, and more trustworthy data for making decisions on allocation of resources, and will be providing itself, parents, pupils, press, and the public with more meaningful reports on achievement in California's public school.

Footnotes

a. Lott, Winsor A. Competency testing in New York State. Paper presented at the American Education Research Association's Annual Meeting. Boston, Massachusetts, April 7-11, 1980, ERIC Ed 194-151.

b. The first volume is called Testing Achievement in Basic Skills: Volume I: Executive Summary and Research Reports. It is part of the final report to the California Policy Seminar, Institute of Governmental Services (Mr. John Cummins, Director), University of California, Berkeley, California, 1982.

Section II

Overview of the Testing Program

Every administrative unit must collect information for making decisions appropriate to its own level of authority. To provide this information, students take tests from the time they enter school until they leave it. Each administrative unit has its own tests. Under its interpretation of a mandate from the legislature, the State Department of Education requires that all students in grades 1, 3, 6, and 12 participate each year in the California Assessment Program (California Education Codes 60601, 60602, and 60640). The state also requires school districts to administer competency tests that students must pass to receive a high school diploma. School districts involved in federal government programs must also administer and report standardized test results to qualify for federal funds. Some school districts administer additional standardized tests to satisfy their own requirements for information. This information may be used to evaluate student progress; determine efficacy of particular educational programs; make decisions on the allocation of resources, such as placement of students in special programs, including honors courses; satisfy requirements for admission to institutions of higher education; and compare achievement against national norms. Students who want to enter the California State University and Colleges or the University of California must take the Scholastic Aptitude Test; for

the University of California they must also take three College Entrance Examination Board's achievement tests. Finally, local schools and their faculties test students throughout the academic year on teacher-constructed and commercially-made tests related to classroom instructional materials, such as progress tests in basal readers. Thus, the state, schools, and teachers collect a large amount of data on student achievement in basic skills and in other educational objectives.

Contrasting Uses of Educational Testing

However, the information gathered by tests of achievement in basic skills changes from attainment at the state and district level of comparative data on general educational achievement to determination at the local school or schoolroom level of student achievement on particular instructional materials. The State Department of Education aggregates testing data for comparing achievement in California schools with achievement in other states, one California school district with another, and each school with itself over time. It interprets these data according to expectancies for achievement at the school district, school, and school grade levels. We will discuss criteria for these expectancies later in this paper.

But, the state testing program makes no attempt to collect nor supply achievement data on individual students nor does the State's

Educational Code allow it to do so (Education Code 60607, 60641). The Education Code (60600, 60601, 60604) only allows the state testing program to determine the level of performance in basic skills at the state, school district, schools within a district, and grade levels within a school. These same provisions of the Education Code enable the State to assess a broad range of objectives. For example, its third grade test encompasses 1020 items. These items are divided into 30 modules consisting of only 34 items per module. Each student takes only 13 language, 12 mathematics, and 9 reading items. In effect, the test uses the scores from modules of 30 students to attain data on all 1020 items. This type of assessment, known as matrix sampling, is permitted by the Educational Code (60603, 60640). While the procedure increases the scope of objectives covered and minimizes the testing time per student, its trade-off is that the State cannot report individual test scores because no student actually takes a complete test.

Whenever school districts need individual test scores on standardized tests they cannot use state test data; they must administer commercially published standardized tests. Each school district then has to pay all the costs for its commercial test and students must take both the State and the school district's tests. We shall discuss this problem again when we evaluate the State testing program in the next section.

Although classroom teachers use information from the district-

administered standardized tests (see Section IV below on school district tests), they also need to construct and administer their own tests. The teacher needs this information to (a) assess students' attainment of specific educational objectives, (b) evaluate the effectiveness of instruction, (c) make reasonable decisions for subsequent classroom instruction, and (d) provide students and their parents with knowledge of progress in achievement of basic skills. A type of test for this purpose is popularly known as a criterion-referenced test, but should be more aptly called an objective-referenced test with a pre-determined criterion for passing the objective(s) being assessed. Frequently teachers set the criterion level for passing the test between 70 to 85 percent of the items on the test. Thus, there are two types of tests for assessing achievement of basic skills: (1) standardized, norm-referenced tests and (2) criterion-referenced tests. (A detailed explanation of norm-referenced and criterion referenced tests and the technical criteria for evaluating them appear in the Appendix.)

Standardized, norm referenced tests are used at the state and school district levels. Criterion-referenced tests more often occur at the classroom level. In describing and critically evaluating the tests at each level, we shall answer these questions:

1. What tests are students required to take and how well do these tests meet requirements for reliable and valid measuring instruments?

2. What purpose(s) do the tests serve?
3. What use is made of the information gathered from the tests?
4. What tests should be administered for which purpose and how should the information collected by these tests be used?
5. What policy recommendations are needed in order to make the current testing program consistent with its intended purposes and uses?

Assessment Assumptions

Our evaluation of tests and testing is based upon the following set of assumptions:

1. Goals of Education
   a. Because tests highlight the goals of education they should assess all the goals of education, including not only basic skills and minimal competencies but also adequacy of preparation of students bound for higher education.
   b. The goals of education should be determined by the public in consultation with professional educators. Because different areas of the state may have particular needs, local communities, through their elected boards, should specify and assess local goals.

2. What purpose(s) do the tests serve?
3. What use is made of the information gathered from the tests?
4. What tests should be administered for which purpose and how should the information collected by these tests be used?
5. What policy recommendations are needed in order to make the current testing program consistent with its intended purposes and uses?

Assessment Assumptions

Our evaluation of tests and testing is based upon the following set of assumptions:

1. Goals of Education
   a. Because tests highlight the goals of education they should assess all the goals of education, including not only basic skills and minimal competencies but also adequacy of preparation of students bound for higher education.
   b. The general goals of education should be determined by society and professional educators. Because different areas of the state may have particular needs, local communities, through their elected boards and their professional staff, should specify and assess local goals.

2. Control Over Tests

   a. State vs. local control. The states should exercise central control over those goals of education common to students throughout the state. Local boards and professional educators should have control over specific goals for their districts.

   b. Public vs. Bureaucratic control. The public should have ultimate control over all testing.

3. Usefulness of Test Information

   a. Tests should provide information that is useful for making decisions about educational achievement.

   b. Information from any test administered in school should have consequences for education at some level. That is, state and school district testers should not only collect test data, but in fact should use the data at some administrative level, and should communicate the results to students and their parents.

   c. Each administrative level has to collect or aggregate data at its own level in order to make decisions appropriate to that level. For example, to determine California trends, the state must collect information on some measuring instrument from all school districts. Furthermore, information that is useful for making decisions at one administrative level may not be useful at another level. For example, norm-referenced tests designed to provide statewide comparative information

are not as useful as criterion referenced tests for making classroom instructional decisions. To obtain useful instructional information, the tests should measure what the teacher has taught. Moreover, the objectives and content of such tests should be teachable and learnable under adequate conditions, which include appropriate sequencing of content, clear directions, and adequate learning time for all students within the normal range of ability to learn.

4. Test Bias

Tests should not have a bias against any group. If a test cannot be bias-free, test makers should either balance its bias or explicitly state for whom the test is appropriate. For example, the Stanford-Binet Test of Intelligence balanced items favorable to females with items favorable to males.

5. Purposes for Administering Tests

a. Instructional decision making. Do these particular students need to learn something? To find out, administer a criterion-referenced pretest that assesses what will be taught. Did these students learn what was taught? Administer another form of the same test after instruction.

b. Selection. Which students are ready to

read, or take algebra, or qualify for a special program? Administer a test that validly differentiates students who meet at least the minimal requirements for the course or the special program. Such a test may be a locally constructed test, designed by the school district specifically for selecting students within a school for a particular instructional program. However, a standardized test with national norms may be the most appropriate test to use for selection of students for college programs which draw students from various high schools and states.

c. Evaluation of instructional programs. Tests provide only one source of data for evaluating a particular instructional program. If the program is of short duration in which only particular students could attain limited instructional objectives, then a criterion-referenced test is likely to be the most appropriate test to evaluate achievement on the program because teachers can tailor-make such a test to have high content and instructional validity. That is, the test should assess what students were expected to learn and what the teacher actually taught. A standardized test, which samples a larger amount of content and may involve more educational processes, is not likely to be as sensitive to short range educational gains. Hence, in determining

gains in achievement in any educational program, teachers should construct criterion-refrenced tests that assess actual classroom instruction and use the information from these tests to complement the data gleaned from a standardized test. In general, tests are a necessary component of any evaluation program but they should be supplemented with other kinds of information, such as observations, interviews, rating scales, and descriptions of instructional method, curriculum, and type of educational content in the program.

d. Allocation of resources. Test data plus other factors such as goals and needs of society, including equality of educational opportunity, national defense, and general welfare enter into determination of resource allocation.

6. Interpretation of Test Scores

a. Cause and effect. Tests do not determine cause-and-effect relationships, but only level of achievement. Further investigation is necessary to determine causal factors for high or low achievement. For this purpose, the legislature authorized the State Board of Education to undertake studies and report on unusually successful practices and factors which appear to be responsible for these practices (Education Code 60601). Among the factors are expectancy criteria.

b. Expectancy Criteria. Students' test performance should be evaluated in relation to criteria on how well students are expected to perform. However, expectancy criteria are not infallible, nor should they be used to act as self-fulfilling prophecies to justify low achievement. Indeed, education is an enterprise in which teachers intervene to ameliorate or improve academic performance. Under some conditions, intervention can result in achievement that is inconsistent with expectations. For example, reading readiness scores at the beginning of grade one usually correlate highly with reading achievement at the end of grade one, but research demonstrated as long ago as 1936 that supplementary instruction provided during grade one drastically reduced the expected correlation.[23] Hence, expectancy criteria should not be used to justify low scores or to maintain the status quo. They should be continuously scrutinized critically to determine whether they are unduly biasing test interpretation. In general, we shall assume all expectancy criteria are fallible and only provide a tentative estimate of expected achievement.

7. Reporting Test Results to the Public

a. As considerable information is needed to appropriately interpret achievement, test scores should be released

to consumers only if accompanied by supplemental data necessary to evaluate the scores. For example, information about a school district's achievement should not be released without a description of the district's expectations, educational plans, and programs for improving student performance. Likewise, to be fair, legislators, newspapers, and the general public should only compare school districts that have comparable characteristics in staff, school facilities, student bodies, and educational programs.

b. Information on achievement which the state, school districts, and school teachers give to the public should be readily understandable to the average citizen. For example, the National Assessment of Educational Progress (NAEP) reports national data by stating the percentage of students in a given grade level who answered particular questions correctly, such as "in 1974-75, 66 percent of nine-year olds could correctly answer a question on information that is directly stated in a passage."[24] The passage and the question are shown in Figure 1. Thus, whenever it is reporting numbers, NAEP specifies

Read the story and answer the question on the next page.

My name is Gregory Gotrocks, and I live in Peoria, Illinois. I sell tractors. In June 1952, the Gotrocks Tractor Company (my dad happens to be the president) sent me to Nepal-Tibet to check on our sales office there.

Business was slow and I had a lot of time to kill. I decided to see Mt. Everest so that I could tell everyone back in Peoria that I had seen it.

It was beautiful. I was spellbound. I simply had to see what the view looked like from the top. So I started up the northwest slope. Everyone knows that this is the best route to take. It took me three long hours to reach the top, but the climb was well worth it.

Why was Gregory Gotrocks sent to Nepal-Tibet?

- ⬭ To climb Mt. Everest
- ⬭ To get away from his father
- ⬭ To make his friends in Peoria jealous
- ⬬ To check on his company's sales office
- ⬭ I don't know.

EXHIBIT 1. Percentages of Correct Responses for Literal Comprehension Item 1, "Gregory Gotrocks" [a]

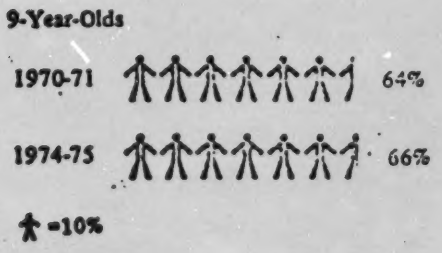


Fig. 1. Example of National Assessment of Education's Report of Test Results

[a] Taken from Reading in America: A Perspective on Two Assessments. Reading Report No. 06-R-01. Denver, Colorado: National Assessment of Educational Progress, October, 1976, pp. 5-6.

the corresponding content or processes on which the numbers have been derived. In other words, if we want the public to understand what students at given ages or grade levels know and can do, we should follow NAEP's lead and only report content-bound data.

Although the California Assessment Program reported achievement in this way in 1979, the California newspapers have not yet publicized its report. School districts and school teachers should also communicate information with numbers tied to content. For example, if a teacher reports a student scored at the 85th percentile on a reading test, the teacher should also state the most difficult passage for which the student could read and answer 70 percent of the questions correctly. Then, a student and his or her parents would know not only how well the student can perform on a test in comparison with a peer group, but also what the student knows and can do on this test. Similar kinds of newspaper reports can be done for data generated by tests administered by local schools and school districts. In other words, instead of reporting only percentile ranks or stanine scores for schools and school districts, newspapers and school personnel should report all test information as NAEP does with content related to test data.

8. Cost effectiveness

Like all expenditure of funds, testing should be cost effective. We shall use two criteria for judging cost-effectiveness:

a. Data are gathered in ways that are most economical and consistent with their intended purposes.

b. Only those data are collected which can be used at some educational level for a specified purpose.

Next, we will describe testing at each level and use our assessment assumptions to evaluate them.

Section II

State Testing: The California Assessment Program

The State of California has gone through three phases in formal achievement testing. In the first phase, begun in 1961, State law required school districts for the first time to test student achievement. Interpreting the law, the State Board of Education selected grades 5, 8, and 11 for testing and allowed school districts to select their own tests from an approved list. The school districts then reported the results to the State Department of Education. But differences among the tests made it difficult for the Department to make comparisons among districts and between the State of California and other states.[3]

Starting in 1965, the State Board of Education ushered in the second phase by adopting one standardized reading test, the Stanford Achievement Test (Reading), for statewide use. The uniform test made it easy to compare the State of California with other states, school districts with each other, schools within a district, and even teachers and students. Newspapers immediately published the percentile rankings for school districts and for schools within their local districts. But school districts, especially those with low rankings, complained vociferously that the test was not valid for their curricula. After three years with the Stanford test, the Board adopted another test, the Cooperative Primary Reading Test.[3]

However, the testing program was still not satisfactory. School districts argued that comparisons based on percentile rankings

alone were unfair. Variations in district scores were so small that a few points difference in raw scores led to a large change in percentile rankings.[2,21] Moreover, the new test still did not solve the curricular validity problem. School districts, especially those with relatively low scores, protested that the test did not reflect their curricula nor what they were teaching in their schools.

The State Board of Education recognized that any standardized test could assess only a limited set of objectives in the typical period of time allotted to testing. Indeed, if a standardized test did assess a comprehensive set of objectives, it would take so much student time that schools would not purchase and use it; hence, commercial test publishers would not even attempt to construct such a test. The Evaluation section of the State Department of Education also noted that normative samples varied from test to test; consequently interpretations of student, school district, and state performance would fluctuate up and down from a change in tests alone! Furthermore, it observed that standardized tests and test norms were updated only after long intervals of time.[3] These complaints about commercially published standardized tests set the stage for the third and current phase of formal achievement testing.

In 1969 the Assembly Education Committee called educators to a meeting to discuss the testing program. The result was passage of the California School Testing Act of 1969, which revised the 1961 law on achievement testing (Education Code 60600). Assembly Bill 665 (1972), together with later revisions, made a number of significant changes in state testing: Testing in grades 1, 2, 3, 6, and 12 was retained;

however, later legislation (1978) dropped grade 2 and expanded the grade 3 test to include reading, writing, and arithmetic. Instead of a reading test at grade one, an Entry Level Test was adopted (Education Code 60640). Scholastic aptitude testing (group or individual intelligence testing) was dropped from all grades except for placement in special educational programs for mentally gifted minors and for research approved by the Superintendent of Public Instruction (Education Code 60614). The most fundamental change was permission for the state to design and construct its own tests and to pay for all costs of testing and scoring (Education Code 60604, 60740)[3].

The State Department of Education exercised its option under the Education Code to develop its own test. Its California Assessment Program under the direction of its Department of Evaluation undertook the development of the CAP tests, known as the Survey of Basic Skills (SBS). The intent of CAP is to collect achievement information on basic skills. To do so, CAP constructs its tests on the basis of the philosophy and principles found in the state-adopted Curricula Frameworks and then uses test items to assess each objective (Education Code 60601). It established advisory committees to review basic subject material at each grade level, formulate objectives for testing, select test items, evaluate the data on the field testing of the items, and approve the final form of the test. For the first tests, CAP purchased test items from commercial test publishers following the selection advice of its four Statewide Advisory Committees in reading, written expression, spelling and mathematics. Table 1 contains a list of CAP tests' names and subtests at each grade level.

Table 1

Tests in the California Assessment Program[a]

Grade Level	Test and Subtests	Items	Skill Area Results Reported in District and School Level Reports (Number of Items in Each Skill Area)
1	Entry Level	35	Total Score: A composite of Immediate Recall (5), Letter Recognition (6), Visual Discrimination (6), Language Development (12)
3	Survey of Basic Skills		
	Reading	270	Word Identification (60): Phonics (30, Structural Analysis (30) Vocabulary (30): Recognizing Word Meanings (16), Using Context (14) Comprehension (150): Literal (73), Inferential (77) Study Locational (30): Alphabetizing (15), Table of Contents (15)
	Written Language	390	Word Forms (66), Standard English Usage (60), Language Choice (3), Sentence Recognition (75), Paragraphs (30), Capitalization (3), Punctuation (30), Spelling (69)
	Mathematics	360	Counting and Place Value (45), Operations (155), Nature of Numbers (45), Geometry (30), Measurement (40), Patterns and Graphs (30), Analysis and Models (15)

Table 1 (Cont.)

Grade Level	Test and Subtests	Items	Skill Area Results Reported in District and School Level Reports (Number of Items in Each Skill Area)
6	Survey of Basic Skills	1,240	1,240 items are divided into 40 unique forms, 31 items per form. Each student takes one form consisting of 10 reading items, 9 written language forms, and 12 mathematics items
	Reading (54 skill areas in 9 categories) Each form has one reading passage. Passages fall into three categories: literature, science, and social studies. NOTE: Content categories represent reclassification of items in skill areas.	418	I Vocabulary (70): Affixes and Roots (16), Word Meanings in Context (37), Multiple Meaning Words (17) II Comprehension (330): Literal (62), Inferential (127), Interpretive (79), Critical/ Applicative (62) III III Study-locational Skills (30)
	Written Expression (39 skill areas in 9 categories. First six of nine items on each form is administered orally)	342	I Writing Process Skills (172): Judging Writing (22), Paragraphs (40), Sentence Combining (50), Sentence Recognition (30), Language Choices (30) II Supportive Skills (160): English Usage (50), Word Forms (32), Punctuation (28), Spelling (50)
	Mathematics (50 skill areas in 9 categories)	480	I Counting Numeration and Place Value (40) II Numbers and Properties (50) III Operations (145) IV Expressions, Equations, and Coordinate Graphs (42) V Geometry (40) VI Measurement (58) VII Probability and Statistics (23) VIII Tables, Graphs, and Applications (30) IX Problem Solving (52)

Table 1 (Cont.)

Grade Level	Test and Subtests	Items	Skill Area Results Reported in District and School Level Reports (Number of Items in Each Skill Area)
12	Survey of Basic Skills	558	
	Reading	144	Vocabulary (31); Comprehension (97)--Literal (47), Interpretive/ Critical (50); Study-locational (13). (Three items not scored.)
	Written Expression	144	Sentence recognition (20), Sentence Manipulation (12), Capitalization and Punctuation (28), Paragraphs (26), Word Forms (24), Language Choices (32). (Two items not scored.)
	Spelling	72	(No breakdown into skill areas)
	Mathematics	198	Arithmetic (98)--Number Concepts (28), Whole Numbers (22), Fractions (26), Decimals (22); Algebra (32); Geometry (24); Measurement (30); Probability and Statistics (14)

[a]Information for this table comes from the Technical Report, California Assessment Program, 1977, p. 7, Survey of Basic Skills: Grade 3, Rationale and Content, 1980, and Survey of Basic Skills: Grade 6, Rationale and Content, 1982.

To reduce testing time, yet cover all the items and objectives, CAP initiated multiple matrix sampling in 1973-74 (Education Code 60603, 60640). It reduced the number of items each student takes to only 25 to 35 items. The total test time is now only 20 to 40 minutes per student.

CAP scores and reports the test results to the legislature and to local school districts (Education Code 60601). However, CAP is not allowed to report individual test scores (Education Code 60607). The report for first grade consists of a school's and district's average scores for the Entry Level Tests, which were designed to assess the entry level of skills for groups of students and serve as a predictor for third grade achievement. The test is not to be used for individual screening, diagnosis, placement, or any decisions affecting individual students (Education Code 60641). Therefore, scores are reported only by school and district which precludes any attempt to use them for evaluating schools with classrooms per grade or individual pupils.

At the other grade levels, CAP reports percent of items answered correctly for reading, writing, and arithmetic at grade 3 (CAP added the basic skills of writing and arithmetic in the 1980 revision of the grade 3 test), three skill areas at grade 6, and four skill areas at grade 12 (Education Code 60602) In 1982, when CAP revised the sixth grade test, it reported achievement in the content areas of social studies, science, and literature by the simple procedure of reclassifying item scores within these content categories.

Teachers supply information to the CAP staff on four student characteristics (Education Code 60601, 60660). The information includes the socioeconomic status of the student's family (parental occupation and educational levels), whether the family is receiving aid for a dependent child (AFDC), whether the child has changed schools (mobility), and English language fluency (FES = Fluent, LES = Limited, NES = Non-English). The first grade Entry Level Test is an index for third grade achievement, a third grade achievement index is used for sixth grade, and a sixth grade achievement index helps predict twelfth grade achievement. From these data the CAP staff, using statistical techniques, estimates an expected level of achievement for school districts with similar student characteristics (Education Code 60640, 60660). An assumption underlying the use of the achievement indices in the expectancy estimate is that the students in the higher grades are comparable to those in the lower grades. Hence, the index accounts for previous level of achievement, which is then used to predict expected level of achievement and weights the expectancy level two to three times greater than the other criteria in the prediction formula. The criteria and weighting used in the expectancy formula can have either adverse or favorable effects in evaluating a school's achievement. For example, if a district is in transition to higher achieving students in the lower grades, the prediction equation will tend to overestimate the expectancy level in the higher grades, and, of course, vice-versa. Table 2 shows the actual expectancy criteria used for each grade level in the assessment program.[3]

Table 2[a]

Summary of Background Factors Used for Computing Comparison Score Bands for 1974-75 and 1975-76

Test	1974-75	1975-76	Source
Reading Test: Second and Third Grades	Entry Level Test Score	Entry Level Test Score	The score from the test developed by the Department.
	Socioeconomic Index	Socioeconomic Index	Teacher recorded for each child occupation of parents in one of the five categories.
	Percent Bilingual	Percent Bilingual	Teacher classified each child in one of the four bilingual categories.
	Mobility	Mobility	Teacher indicated whether the child has been continuously enrolled.
		Percent AFDC	District provided data by school's attendance area.
Survey of Basic Skills: Grade 6	Grade 3 Achievement Index	Grade 3 Achievement Index	Immediate prior year scores from the statewide testing program. Tests were not the same across two years.
	Socioeconomic Index		School principal gave percentages of students in the four occupational categories.
	Parental Education Index		School principal gave percentages of students in the four educational categories.
		Percent Bilingual	Teacher categorized each child in one of the four bilingual categories.
		Percent AFDC	District provided data by school's attendance area.
Survey of Basic Skills: Grade 12	Grade 6 Achievement Index	Grade 6 Achievement Index	Immediate prior year score from the statewide testing program. Some test items were not the same across the two years.
	Socioeconomic Index		School principal indicated percentages of students in the four occupational categories.
	Parental Education Index		School principal indicated percentages of students in the four educational categories.
		Percent AFDC	District provided data by school's attendance area.

[a]This table was reproduced from California Assessment Program[3]

Since any estimate has some chance variation in it, the CAP staff also constructs an expectancy band which consists of the estimated expectancy score plus and minus the chance variation. The expectancy band which represents the rankings of the middle 50 percent of schools having similar backgrounds, indicates what each school district's achievement should be relative to other California school districts. But until 1980, CAP did not differentiate expectancy bands according to school district size, which resulted in some inequitable comparisons. CAP also reports year-to-year comparison of district mean scores, except of course when it revised a test, which it does every 4 to 6 years. However, under its recently adapted statistical model, a modified Rausch model, and a new scaling procedure, the State Department claims it can make year-to-year comparisons in spite of changes in the test.[5]

CAP breaks down achievement according to four major student characteristics within each school for each grade tested. In the 1980 test revision, schools got breakdowns of scores and expectancy bands for objectives within tests, which make the test results more useful to the school, and scaled scores which have nearly equal intervals and allow for direct comparisons across tests, between schools, and across years. CAP also provides tables for transforming the scaled scores into percentiles and into percent correct on the tests.

Although CAP has its uses, like all tests, it also has its limitations. We shall state both its uses and limitations.

CAP Uses

At the state level, the Education Code (Section 60601) requires the State Department of Education to use achievement test results for evaluation of programs, allocation of resources, and identification of unusually successful practices and the factors which appear to be responsible for attainment of high achievement. Alexander Law, Chief of the Office of Program Evaluation and Research, California State Department of Education, claiming that CAP does what it is supposed to do, reported that CAP data have served as the source of information for evaluating "virtually all major categorical programs and reforms efforts, including Early Childhood Education (ECE), Miller-Unruh Basic Reading, and programs for educationally handicapped youths, bilingual students, the educationally handicapped, and mentally gifted minors." Furthermore, he pointed out that CAP's evaluation of math programs showed that math achievement which had been dropping leveled off and began to rise after the math framework and textbooks were shifted away from the "new math" approach. He also credited state testing with providing early warnings about a drop in writing ability. Law noted that state data were also useful in identifying schools for a "School Effectiveness Study" and for similar studies. He concedes that uses of CAP data are "difficult to achieve" at the school level. However, he claims a survey of schools found that 55 percent of the school districts which responded to the survey had made program changes to correct weaknesses identified by CAP results.[9]

CAP annually publishes documents that supply technical information for evaluating its tests, except for some technical information on the items.

The information which CAP supplies in its Technical Report, Annual Report on Student Achievement in California Schools, Interpretive Supplement accompanying school and district reports, its Profile of School District Performance accompanied by A Guide to Interpretation and Survey of Basic Skills: Rationale and Content provides a fairly good basis for evaluating the CAP tests.[3,4,5] Although they have limitations similar to other standardized tests, such as a multiple choice format for assessing written expression, nevertheless, according to technical standards for judging a test, the CAP tests appear to be well-constructed, standardized tests with norms that are exhaustive for California schools and school districts.

Nevertheless the CAP tests have their limitations. These limitations are in (a) the construction of the tests, (b) interpretation and misinterpretation of test results, (c) validity of the data for program diagnostic information, (d) effectiveness of CAP, (e) CAP omissions and (f) issues of public policy on goals and control of schools.

CAP Limitations

Test construction

Extensive time, money, and effort have gone into the construction and execution of CAP. The CAP staff has attempted to design instruments that are relevant to California schools and has probably succeeded in selecting objectives and writing test items that fit the

California Framework Statewide Curriculum (Reading, Mathematics, and English language). Indeed, CAP uses a comprehensive set of objectives to cover the multiple programs used in reading, spelling, language, and mathematics instruction in all California public schools. However, CAP still equates its current tests with commercially published standardized tests and must assume that the CAP tests are measuring the same objectives as the standardized tests in order for CAP to arrive at comparisons with national norms.

Although CAP provides a Technical Report which contains almost all the information needed for evaluating the CAP tests, CAP has not made available to school districts information about which of its test items are assessing which objective and the percent of students who pass each item. Nor has CAP distributed information on relevant materials for improving student performance on the CAP objectives. Without such technical and user information, known as a "Test Atlas," it is difficult for a school district to decide where and how to modify its programs. Moreover, CAP has not made information available on its test items' characteristics and scoring procedures. Such information on test items is essential for judging the validity of the items, their grade placement, their relationship to the curriculum, and their fit to the statistical model for scaling a test. The scaled scores permit comparisons of scores across and within basic skills achievement, among schools, and within schools over the years. However, without complete information on the construction, development, technical characteristics, statistical procedures, administration, interpretation of scores, and factors that affect test scores and their

reliability and validity, which, according to the American Psychological Association's Standards for Educational and Psychological Tests, constitute the essential requirements for evaluating any test, we cannot adequately evaluate the CAP tests.

Interpretation of Results

CAP can readily determine and report trends within the state. Law reported in 1978 that reading achievement in the primary grades consistently improved since testing went into effect in 1966. In 1978 the California average was "well above the national average." But he noted that at grade 12 scores have consistently gone down, declining the most in writing and the least in arithmetic. More specifically, the median California pupil performance in third grade for 1966-67 through 1979-80 increased in reading achievement from the 34th percentile on the Stanford Reading Test to the 58th on the CAP tests, called the Survey of Basic Skills (SBS), equated to the norms on the Comprehensive Test of Basic Skills (CTBS). Over the same time period, the sixth grade has gone from the 48th percentile rank on the CTBS to the 56th percentile rank for reading on SBS (equated to CTBS norms), 43rd to 53rd on language, and 47th to 55th on mathematics. Concomitantly, the 12th grade has dropped from 52nd, 42nd, and 48th percentile ranks in 1969-70 on the Iowa Test of Educational Development (ITED) to 41st, 34th, and 44th for reading, language and mathematics, respectively, on the Survey of Basic Skills (equated to ITED norms).[3,4,5]

However, the CAP and standardized tests are not based on the same content or processes. CAP derives a total score by summing all the

elements or subtests for a basic skill; for example, phonics and reference skills are added to comprehension skills to get a total score for reading at grade 3. (See Table 1 for a list of the elements at each grade level.) Then CAP compares this total score with standardized test scores for "reading," but the standardized test scores may be based only on the comprehension components!

Also, comparisons with "national norms," which are required by the Education Code (Section 60640), are questionable. If the same tests had been given to at least a representative sample of students in each state at the same grade levels tested in California, the results of this testing could have led to the construction of norms with equivalent percentile rankings for mean grade level scores. But no such direct testing has been done. Instead, the percentile scores reported to the legislature and the public are based upon a host of statistical assumptions and manipulations. The process of deriving equivalent percentile rankings is complex since any one student takes only 3 percent of the CAP test items. Then the CAP test results have to be equated with standardized tests in order to use the standardized tests' norms. The outcome of this norm-equating process can vary considerably according to the standardized test used. We can illustrate the variation in national percentile ranks by using the data in the annual CAP report for the years 1965-66 through 1978-79.[3,4,5] The marked increase in reading performance between 1970-71 and 1971-72 was partially due to a change from the Stanford Reading Test to the Cooperative Primary Reading Test. The norms on the Cooperative Test resulted in higher equivalent scores than do those on the Stanford

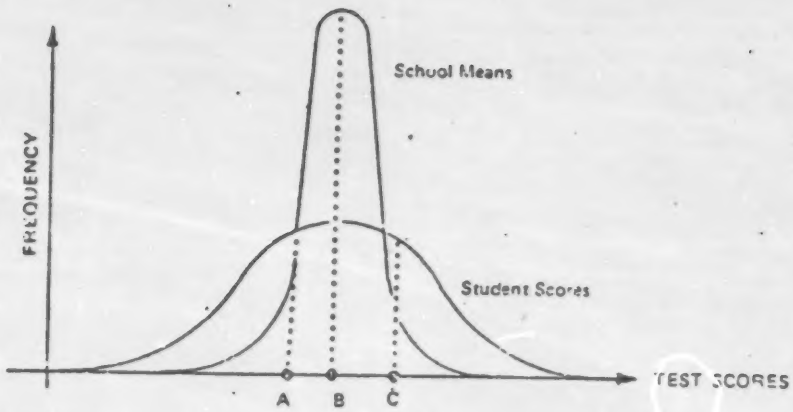
test. Ironically one of CAP's justification for constructing its own test was to avoid use of such outdated commercially published test norms. Although CAP has up-to-date norms for comparisons within California on its own tests, when it equates its tests with national norms for interstate comparisons, then CAP short-circuits its justification for constructing its own test!

Moreover, without knowledge of the distribution of state scores, we have no idea of how many raw score points can lead to a change in California's percentile rankings when CAP compares California with the other states. If the mean scores of the 50 states are close together, as we suspect they would be, a very small change in raw score points can result in a very large change in percentile ranks. For example, Law reported that in reading achievement third graders were at the 58th percentile, but 12th graders had dropped to the 34th percentile on national norms.[9] How many raw score points these variations from the national average represent was not in his report. Nor did he give any explanation for this drop. However, Willard Wirtz, former Secretary of Labor, who chaired an advisory panel on the decline in scholastic aptitude test scores, found that "about half the score decline, perhaps more, is a result of our deciding in the 1960s to keep more young people--which meant a larger percentage of lower scores--in school longer; and the rest of the decline apparently reflects various bad habits that crept into the schools and society, including everything from social promotion in the schools to social changes in the home...."[25]

The problem in interpreting scores at the state level also occurs with comparisons made among school districts in California. The mean achievement scores for the school districts are bunched close to each other, clustering around the 50th percentile. As a result, a small change in the district's mean score can produce a very large change in the percentile rank for the district; for example, as shown in Figure 2, a change in percent correct on CAP from 58 to 68.7 can change a school's percentile ranking from 19 to 91. The figure also shows that students' percentile ranks are different from those for schools; the same change in percent correct would change a student's percentile rank from 41 to 74 on the commercial publisher's norm.

This narrow distribution of scores can produce severe interpretation problems for educators, legislators, and the public. A southern California school board member's comments illustrate this point:

> When the students in my district took the grade 12 written expression test a few years ago, they answered 57 percent of the test items correctly. This past spring the students in grade 12 answered 63 percent correctly, for a gain of six percentage points in two years. This means district seniors were doing better, right? Wrong! Or, maybe wrong, or it depends on how much one knows about testing, or on how bright the local reporter



	A	B	C
CAP Percent Correct	58.0	62.0	68.7
Publisher's Student Percentile Rank	41	52	74
State School Percentile Rank	19	41	91

Fig. 2. Comparison of student and school percentile ranks[a]

[a]From the California Assessment Program, Interpretive Supplement to the Report on Survey of Basic Skills: 1979-80.

is, or on how the newspaper reports the results! That six-percent increase in the number of items answered correctly translated into a normative score change from a percentile rank of 76 in 1975 to 68 in 1977. (Percent correct went up six points.) During those same comparison years at sixth-grade level, our schools on the average scored seven percent higher in 1977, but the norm score reported to us by the State Department dropped one point.[2]

Validity of CAP for program diagnostic information

According to CAP's Annual Report (1978-79), CAP tests provide "program-diagnostic information for local and state planning." Given the current design of CAP tests we must seriously question the validity of this claim and the feasibility of even achieving the purpose. The absolute minimum level of meaningful diagnosis cannot be at the individual level because individual students do not take a complete test. Nor can it be at a classroom level unless a school has only one classroom per grade. Why? Because the smallest unit for which CAP returns information is only for the entire grade within a school. Consequently program-diagnostic information is only relevant for the entire grade. If only some students within a grade participate in a special program, a school cannot use CAP data to evaluate the effectiveness of the program because it cannot divide up scores for students within the grade. The school can only determine the effect of its overall programs of basic skills for the total grade and for the four characteristics of students for whom the CAP scores are

"broken out," that is, reported separately, such as socioeconomic status, males vs. females, mobile vs. non-mobile students, and English fluency. Moreover, the school does not get any subtest scores which are necessary for determining strengths or weaknesses in a program until the third grade revision in 1980. But CAP does not provide individual scores. So the school cannot determine which students are strong or weak in particular subskill areas. The result is that program information without student scores is like knowing the strength and weaknesses of medicines, but not knowing which patient should get which medicine! Also subtest scores come with their own expectancy bands. Although such information implies school districts ought to improve in weak subtest areas, CAP points out that school districts need not improve in each subtest.[5] But how else can a district improve its scores on CAP tests?

CAP Effectiveness

At best CAP provides only highly generalized suggestions for state, school district, and local school educators for establishing relationships between program or instructional effectiveness and achievement performance on the CAP. We can best view CAP results as a first step screening to identify variations and trends in achievement levels. CAP's 1979 survey of elementary school principals on the usefulness of CAP test results supports this conclusion. Principals (74 percent) found the CAP tests were useful for comparing results across years, but only 24 percent reported the CAP data were helpful for planning new school programs. What adversely affects CAP's usefulness is its high frequency of test revision. For example, CAP

revision interfered with its reports of trends. Moreover, when it renormed its tests, which it did every year, it meant that each year 50 percent of California school districts would be equal to or below the average.

However, CAP claims that its new scaled scores, first introduced in 1980 with its revised test for grade 3, will solve all these test comparisons, even when the state revises its tests. The scale will not be renormed or adjusted in any way each year; consequently a school can monitor its progress independently of the progress of other schools. Moreover, the scaled scores will make it possible to make direct comparisons among skill and content areas, and between different tests.[5] However, the validity of these claims has not yet been demonstrated.

Up to this point, we have described and evaluated what CAP attempts to do. We now turn to the testing functions that CAP omits.

CAP Omissions

The legislature decided to let the State Department of Education determine whether pupils should be administered the entire test or "only a portion of the test which will be representative of all the test objectives, goals, or categories of items on the entire test" (Education Code 60602). But the Department was forbidden to report the "score or relative position of individual pupils" (Education Code 60607). Consequently, CAP provides no information for determination of individual levels of achievement in basic skills and for diagnosis of these levels.

However, teachers need information on individual students for planning instruction to eliminate achievement weaknesses. Moreover, without individual test scores teachers cannot communicate to parents how their children are performing in achievement of basic skills. To fill this gap in testing functions, the legislature decided that testing for the purpose of obtaining individual scores should best be left to the local school districts and therefore allowed the governing board of each district to adopt its own tests (Education Code 60605).

Schools also need individual test scores for placement and certification of students. They need to know which students have attained prerequisite skills for certain courses, such as algebra, physics, or any subject which have prerequisites that must be fulfilled by students in order to have a reasonable chance of success in the next step in a sequential program. Schools must also certify student achievement not only for parents but also for college and university admissions. Also, federal programs require individual test scores on standardized achievement tests for school districts to qualify for federally-funded programs.

Evaluation of instructional programs are also necessary for a school district's educational program. Although CAP provides group information that is relevant for program evaluation, this information occurs only at grades 3, 6, and 12. For programs in-between these levels, neither the school nor the state can determine whether they are effective or not. For example, a junior high school may be excellent but feeds into a very bad high school. Current CAP test results would provide no knowledge of this variation. We can also cite

another example of difficulties in program evaluation that arise from grade level gaps in the CAP testing program: When he was President of the State Board of Education, Michael Kirst wondered where the drop-off in achievement observed at the 12th grade level had actually occurred; he could not tell because he had data only for grades 6 and 12. However, this problem will be at least partially solved if the legislative analyst's recent recommendation is adopted that CAP should also test in grade 8.[26] But the gaps in the other grade levels will still remain.

Although CAP's adoption of scaled scores solves its renorming problem, it still does not eliminate possible misinterpretations of test results because the scaling process is difficult to comprehend. In an attempt to make CAP test data comprehensible some newspapers are likely to use the tables provided by CAP to translate scaled scores back into percent correct and percentile rankings which will re-introduce the misinterpretation that the new scaled scores tried to correct!

To provide test scores for individual students, school districts may adopt their own standardized tests for assessing achievement in basic skills. However, some school districts do not exercise this option for various reasons, including the expense of testing. Hence teachers in these districts do not get information for diagnosing individual students' strengths and weaknesses nor for determining causes of failure to achieve, parents do not obtain information on their children's progress in basic skills, certification for programs is not available, and school boards have no standardized test data for

evaluating their special programs and instruction at grades other than 3, 6, and 12. Ironically when the testing program began in 1961 and up until 1972, the law and the State Board of Education required school districts to administer standardized tests in grades 5, 8, and 11. With the advent of the Assembly Bill 665 (1972) leading to the establishment of the CAP program, the state dropped this requirement.

Conclusion

Now, to answer the question we posed earlier: To what extent has CAP met its objectives? We shall answer this question from the standpoint of the Legislature, the State Board and the State Department of Education, school districts, and the public.

CAP results provide current and trend indications of achievement in basic skills. While the information has to be evaluated and interpreted carefully, the Legislature and the State Board and the State Department of Education may generally find trend information necessary as a initial step in making decisions on evaluation of programs, allocation of resources, identification of successful school practices, and on research in causal factors underlying achievement in basic skills.

At the school district level, CAP provides data on achievement trends and achievement in percentiles and percentages at grade levels 1, 3, 6, and 12. Scaled scores will begin in 1980 at grade 3 and then in grades 6 and 12 when CAP revises its tests at these grade levels. CAP also helps school districts evaluate their test results by

providing expectancy bands and by reporting results according to group characteristics, including, but only recently, percentage of students in each quarter of the statewide distribution of student scores. At each grade level CAP will also provide a "breakout" of total test scores into their subparts with expectancy bands for each subpart.

The CAP results may have value in formulating first level screening for identifying what appear to be "high and low achieving districts." However, follow-up on this screening is necessary and critical in order to determine which of a multitude of variables influence these high and low achievement levels. Indeed, allocation of funds to districts on the basis of information provided by CAP alone without knowledge of local school district conditions and curricula would be inconsistent with principles of justice and equality of educational opportunity.

Although CAP provides information to the public, its information, like all test data, usually gets to the public in the form of newspaper stories which only confuse rather than aid the public's understanding of achievement progress. The public usually gets scores in percentiles, without the content or skill levels associated with these scores. Content-free scores do not inform the public on what students know and can do in basic skills.

Perhaps newspapers have problems in reporting CAP results because the results are difficult to comprehend. Although CAP's recent adoption of NAEP's content-relevant method of reporting has clarified the meaning of CAP scores, reports such as the one made by State

Superintendent of Schools, Wilson Riles, that California high school seniors achievement in reading has dropped steadily and is now at the 34th percentile on national norms got newspaper over-the-mast, front page headlines, but did not explain anything to the public who wanted to know what this drop means and why it occured.

At present, the State Department of Education bypasses local school districts in the initial release of CAP results to the public. Hence, newspapers do not accompany the results with the local school district's interpretation. Nor do the newspapers relate the results to the school district's plans or programs for developing basic skills and improving test scores on its own standardized tests. Unfortunately, misinterpretation of the test results, particularly low ones, can erode confidence and support of the local community not only for their local schools but also for public education in general. It is therefore imperative that newspapers interpret test scores properly and in relationship to whatever program the school has adopted for instruction in basic skills.

Another major shortcoming of the CAP tests arises from the fact that students take only a three percent sample of test items, which means that an individual's score on any one test is not reliable enough to report individual test results. Moreover, the Education Code does not allow CAP to return scores to individuals, even if they were reliable. This shortcoming is tantamount to getting on a scale for measuring weight but not being able to find out the result. Knowing they will not see their own scores and indeed that

individual scores are not forthcoming, some students are likely to stop before they even get on the scale or if they take the test, they do so only perfunctorily. Thus this shortcoming tends to reduce students' motivation for taking the CAP tests. Hence their scores on the test do not reflect their true achievement; their scores are lower than they should be. Indeed reduction in motivation may explain some of the decrease in test scores at the 12th grade.

Teachers may have similar attitudes when they realize they cannot relate scores to individuals or even find out how their class as a group performed, unless they happen to be in a school where there is only one class and it has more than the minimum size necessary for a grade level report. Hence the teachers are likely to believe the tests are an intrusion on their instruction time. Such attitudes can also reduce the reliability of the tests.

At the state and district levels, the test information may be useful. But as the test approaches the school, classroom, and individual levels its weaknesses increasingly become paramount. This change in the usefulness of the test became apparent as we interviewed a representative sample of educators and legislators and their staff personnel. To summarize this interview data which is reported in detail in an accompanying volume, Testing Achievement in Basic Skills Volume I: Executive Summary and Research Reports, educators at the district and administrative levels were more knowledgeable than teachers about the CAP tests and found the data somewhat useful in

evaluating curriculum but tended to complain that too many skill areas were assessed. The lawmakers and their staffs indicated that it was important to have test data on basic skills, but recognized that the State Department of Education has an inherent conflict of interest in constructing, administering, and interpreting test results that reflect upon its own administration of California public school education. Some lawmakers and their staffs suggested that the conflict can be resolved by contracting with an independent agency for evaluation of basic skills[14]

Policy makers, decision-makers, and ultimately an informed public will also have to decide whether the trade-offs in the CAP tests are worth the costs or whether it would not be more economical to curtail or even eliminate the CAP tests and give, instead, a sample of California students the nationally standardized tests that CAP equate its tests with, and thus avoid the roundabout route CAP now takes to get national percentile rankings. Regardless of the solution to the testing problem itself, a decision will have to be reached on elimination of the State Department's inherent conflict of interest in the testing program, if the test results are to be used at the legislative level for the decision making process in allocating of funds. These issues are not technical issues but public policy issues. We shall discuss public policy issues at greater length and make our recommendations on them in our Executive Summary. Next, we shall review school district testing.

School District Testing

Under the reserve clause of the constitution, control over education has remained with the states. Nevertheless, shortly after the constitution had been ratified the federal government began to influence education with the passage of the Northwest Ordinance in 1789, which set aside land to be used for schools. Federal influence on education increased slowly until some 30 years ago when two events initiated a dramatic increase: (1) a Supreme Court decision and (b) the National Defense Emergency Act.

In 1952 the Supreme Court rejected its 1896 decision, which had sanctioned segregated education. In reversing itself, the Court concluded that "separate but equal facilities" were inherently unequal. The decision not only outlawed de jure segregated education, it also changed the concept of equality of educational opportunity from a criterion of equal input of resources to equal output in achievement. To attain equal output, schools could now provide unequal input. The schools had long been doing so for the education of exceptional children, but the Supreme Court expanded the concept to include minority groups.[27,28] Many issues in education stemming from the 1952 Supreme Court decision have been and are still being resolved in federal courts.

The National Defense Emergency Act, passed in 1958 by Congress, acting under the general welfare provision of the Constitution, increased federal influence in another way. The Act provided federal funds, particularly for training of mathematicians, scientists, and

engineers, to schools complying with federal guidelines. However, evaluation of federally funded programs was not required till passage of the Civil Rights Act of 1967. Under this Act, school districcs now have to use standardized achievement tests to assess programs funded under its various titles.

Although federal influence over education has incresed, the states still predominate in control over education. However, local school districts have vied with the State over control of education. This conflict can be perceived in textbook selection. State textbook adoption began in 1883 to insure that California students would get the best materials available. Free textbook distribution of state adopted texts started in 1912.[29]

However school districts wanted to have control over textbook selection. Senate Bill 1 (1968) eliminated single textbook adoption, and substituted multiple adoption, which consists essentially of a list of textbooks that the curriculum commission certifies as meeting its educational and legal compliance criteria. However school districts can adopt any textbook, but if it is not on the list, the school districts must have the curriculum commission certify it the next year. If the certification is not granted, the school district loses its investment because it can no longer use the textbook. In place of free distribution of textbooks, the legislature now distributes funds on a per pupil basis for the purchase of textbooks, but the amount of money covers purchase at most of only one textbook per student per year.

The state adopted testing program counterbalances the change in textbook selection. Local districts have greater control over input

via selection of teaching materials; the state now monitors the output, achievement on the state-mandated CAP tests of basic skills. However, school districts also monitor achievement by administering their own tests.

Achievement testing began in 1897 in the United States with a spelling test, created by Joseph Rice. A parallel development was taking place in France that also had great impact on testing in the United States. In 1905-1911, Alfred Binet and Henri Simon developed the Binet-Simon Scale for determining intelligence. The test, later revised and known in this country as the Stanford-Binet Test of Intelligence, was originally used to determine by objective means which students were eligible for France's new residential treatment center for mentally retarded children. The test served not only to select children for the institution, but also to protect students who did not belong in the institution from parents and teachers who had their own motives, either benign or malicious, for wanting the children in residential schools. This test was widely used in California until its use was restricted to research and to certifying placement of mentally gifted minors (Education Code 60614).

However, achievement tests have thrived. During the period from 1900 to 1925, achievement scales in spelling, handwriting, arithmetic, and reading were invented. These tests ushered in the scientific movement in education. The purpose of tests in this movement was to develop an adaptive curriculum to prevent student failure and to provide suitable educational programs for students with learning problems. In the era of 1930-70, tests were also used for guidance. More recently with the advent of proficiency testing they became a

determinant for certifying minimal competency necessary for receiving a high school diploma.[13]

Although the scientific movement was subsequently replaced by other emphases, for schools are institutions of society and reflect changes within society, the use of commercially-published tests has developed into a huge industry. They are now used for a wide variety of purposes, in and out of school situations. For example, school districts use them to generate answers to the following questions for their constituents and themselves.[30]

General Public. What is the quality of education in the district, and what evidence is there for equality of educational opportunity among the various schools in the district and for exceptional students, including gifted, handicapped, bilingual, or limited or non-English speaking groups, and its ethnic minorities? Although no consensus exists on evaluation criteria, in practice a school district does not use test scores alone to determine the quality of its educational program, but uses a variety of factors, such as the following:

a. Input: Per pupil expenditures, class size, teacher preparation and training, mobility of students and faculty.
b. Educational Process: General and special programs, including those for exceptional minority students; laboratory and library facilities; and extracurricular activities.
c. Results: Achievement test scores; disciplinary referrals,

including suspensions and expulsions; drop-out rates; college admission rates; scholastic aptitude and College Board Advanced Placement test results; Merit scholars; and student attitudes toward school. But, the two top-priority items agree with the Gallup Poll results: the public wants improvement of basic skills and better discipline.

Parents. Parents want evidence that their children are learning, that they are safe, and that discipline is fair and effective. They also want to know that school is a reasonably pleasant experience and that their children feel positive towards school and towards themselves as students.

School boards. In addition to the above concerns, the school board wants to know how the district's achievement compares with demographic trends and whether it actually exceeds such trends. It is also interested in cost effectiveness and accountability of teachers, principals, and central staff.

State and federal agencies. The state wants achievement data, such as obtained by the California Assessment Program in grades, 1, 3, 6, and 12, as well as attitudinal and self-concept data (AB 65 School Improvement Program). It also requires schools to administer a norm-referenced test for assessing fluent English speaking (FES), limited English speaking (LES), and non-English speaking (NES) for the Compensatory Education Program. In addition, the federal government requires pre-post standardized achievement data for evaluating achievement in categorically funded state projects, such as Title I. The

school district must also collect nontest data to determine target students in selected programs: mobility data, aid for dependent children, and parents' educational and occupational levels. Furthermore, criterion-referenced test results on a school-by-school basis are required each year for the school improvement program, for compensatory education, and for the bilingual program. Finally, the federal government from time to time requires resumés of longitudinal test data.

In June 1980, state competency test requirements went into effect. Each school district had to construct or select its own test and determine its own passing score. Seniors who do not pass the test may be given a certificate of attendance instead of a diploma (Education Code 51217).

Thus, the school district administers both norm-referenced and criterion-referenced tests to answer questions posed by the general public, parents, school board, school administrators, and by the state and federal governments. In addition to trying to satisfy these purposes, the school district may use test results for diagnosing individual students' learning problems, certifying students' competence, and evaluating the performance of district schools, students, and special projects.

Uses and Limitations of Tests and Test Results

Test results have their uses and limitations for answering questions posed at the district level. In this section, we will review the California Assessment Program's uses and limitations from a

district point of view. In doing so, we will unavoidably overlap somewhat with information presented earlier. We shall also examine the role of commercially-published standardized tests in answering district-level questions.

California Assessment Program. CAP reports two types of test results to school districts. One type communicates the district's percentile ranking in comparison with other school districts. However, this type of report is sometimes more detrimental than useful to a school district because the press and the public frequently interpret the percentile ranking as the answer to the question of the quality of education in the district. But this single score, if taken as an index of a school's educational quality, is a vast oversimplication to a complex question.

Those who are not knowledgeable about measurement may misinterpret the CAP score itself. For example, because the distribution of CAP scores is apparently quite narrow, a change in performance of a few raw score points on the test can mean a large percentile ranking change for a school district. Also, since percentile rankings for the districts were rearrayed each year from 1 to 99, half of the districts were always below the 50th percentile. Under this annual renorming process districts could have increased their raw scores from one year to the next, yet could have found they had decreased in percentile rankings because the other school districts had increased more. Hence, this annual renorming of CAP data led to misinterpretations. Although CAP's switch to reporting only scaled scores will tend to

eliminate this problem, the oversimplication inherent in the use of CAP test alone to judge the quality of education in the state or in a school district will still remain. Furthermore, no matter how CAP changes its scaling, it will still need to get a comparison of California's achievement with other states by equating CAP data with publishers' norms on commercially prepared, standardized tests. Ironically, CAP justified construction of its own test because the publishers' norms were so much out of date! Hence, its interpretation of achievement in California will still be susceptible to the limitations of commercial-publishers' norms. If the legislature were only interested in trend data and comparisons of educational achievement in California with other states, it would be simpler and more economical to give a standardized test, instead of CAP, to a sample of California students in the first place.

Some districts would prefer to have the state report only state trends and not try to apply CAP at the district level. For determining state trends, a sampling of schools within a district would be sufficient.[21] It would save the state considerable money and would not use up the expensive instructional time of all students in grades 3, 6, and 12. The state could use CAP or a commercially prepared standardized test for this purpose.

Of somewhat greater value to the school district than the CAP percentile rankings is another type of CAP report. This report shows for each grade level in each school the average percentage correct on the CAP tests for all the students. It also indicates the average percentage correct broken down for males vs. females, mobile students,

socioecnomic status, English fluency (fluent, little, or non-English). However, unless the school has only one classroom per grade, the test results cannot be related to instructional programs in particular classes. Nor are the data useful for program evaluation of grades in between 3, 6, and 12 because CAP does not assess these other grades.

CAP started reporting scaled scores for its grade 3 test when it was revised in 1980. It will do likewise for its 6th and 12th grade tests when they are revised and for its new 8th grade test when it is constructed, probably in 1982. Scaled scores have an arbitrary mean of 250, which has some similarity to the Consumer's Price Index. It has several advantages: unlike percent correct scores or percentile rankings, the user can compare scaled scores from one content area to another ("Is the school better in math or in reading?"); the performance of two districts can be compared because the scaled scores have nearly equal intervals ("How much better is District A than District B on this test?"); scaled scores can be compared across years regardless of test changes or amount of statewide progress ("Is my district's performance in math better this year than last year?"). Although scaled scores provide comparable information and are easy to use, they are very difficult to understand.

Some school districts also have some reservations about the reliability of CAP criteria because teachers code the data and because the code is susceptible to variation in interpretation and to misinterpretation, particularly in determining language fluency, which is one factor that weights the expectancy formula. Since teachers

provide this information on students, some variation may occur in how teachers interpret the criteria for the information categories. If some teachers are lenient and place too many students in the little or non-English speaking category, it will lower the achievement expectancy for their schools, and of course, vice-versa. If a lower expectancy is computed, the school district could get a better achievement performance rating with the same achievement data.

If school districts had complete confidence in CAP, they would still need an Atlas, a list of objectives assessed by CAP and items information related to these objectives, with objectives cross-indexed to the California Framework for Education and to texts in current use in California. With the Atlas, they could diagnose instructional elements that are weak. Nevertheless, some district directors try to relate CAP's data to program plans in each school. In general, they seek to infer a cause-and-effect relationship between school program plans and changes in achievement data. However, as in any research project, test data can only point to a relationship; they do not indicate cause-and-effect. Consequently, schools need to conduct further research to answer the question of what caused the change.

For reports on individual achievement in basic skills, districts must give their own standardized test. Since the districts also have to give the CAP tests, they believe they are involved in duplicate testing.

Given all these limitations of CAP for evaluating schools, programs, and students, research directors often prefer to select and rely on publishers' standardized tests and a criterion-referenced

tests whose contents are clearly described, whose construction is understandable, and whose test specifications can better fit particular program objectives. Without such understanding, it is not surprising to learn that some districts rely not on CAP data, but on their own standardized tests results.

Standardized Test Data. Districts vary in their use of standardized tests. For example, each year during the elementary grades the Riverside Unified School District administers the Stanford Achievement Test Battery. The Alvord Unified School District administers the Comprehensive Tests of Basic Skills at the beginning and end of the school year. The Metropolitan Reading Instructional Tests are also given in Riverside and elsewhere. In Table 3 we describe the Comprehensive Test of Basic Skills and the Iowa Test of Educational Development because through a statistical equating process CAP uses the national norms from these tests for interpreting CAP test results.

In addition to reporting individual school achievement scores to parents, pupils, and the public, Riverside also uses the standardized test results in reading as a kind of management system which it calls RISS (Riverside Instructional Support Service). It groups item-scores by objectives and cross-indexes them with school materials. Teachers can then determine which objectives students need to achieve and can readily locate materials for instructing students in these objectives. This management system, done on the Stanford Achievement Test, is what CAP could do in constructing an Atlas for its tests. However, a problem arises in the use of RISS: Considerable time elapses between test administration and return of

Table 3

Standardized Norm Referenced Achievement Tests for Basic Skills Whose Norms are Used by the California Assessment Program Tests

Comprehensive Tests of Basic Skills (published by CTB/McGraw Hill, Del Monte Research Park, Monterey, California, 93940, 1977).

Form T. (Basic Battery), Level 1, Grades 2.5 to 4.9; Level 3, Grades 6.5 to 8.9 [a]

Reading (46 to 56 minutes)

Test 1. Reading Vocabulary: Select synonym for underlined word in a phrase.

Test 2. Reading Comprehension: Literal recall, rewording, context clues, main idea, descriptive words conclusion, style and structure.

Language (48 to 50 minutes)

Test 3. Spelling: Recall of rule - recognize correctly and incorrectly spelled words; Context clues - decide which pair of homonyms or easily confused words is used correctly in a sentence.

Test 4. Language Mechanics: Punctuation and Capitalization

Test 5. Language Expression: Usage, diction, syntactical relationship, organization.

Mathematics (70 to 75 minutes)

Test 6. Computation: Add, subtract, multiply, divide; raise to a power; extract a root.

Test 7. Concepts and Application: Number systems, geometric relationships, measurement, sets, graphs, math sentences, problem-solving, reasoning.

[a]Revised forms U and V are now in use.

Table 3 (Cont.)

Iowa Tests of Educational Development (published by Science Research Associates, Inc., 259 East Erie St., Chicago, IL 60611, 1962.)

Form X-4, Grades 9-12

Tests	Time	Items
1. Understanding of Basic Social Concepts	55	90
2. General Background in the Social Sciences	60	90
3. (a) Correctness and Appropriateness of Expression	60	84
(b) Spelling (detecting incorrectly spelled word)	65	14
4. Ability to do Quantitative Thinking		
5. Ability to Interpret Reading Materials in the Social Studies	60	53
6. Ability to Interpret Reading Materials in the Natural Sciences	60	80
7. Ability to Interpret Literary Materials	55	80
8. General Vocabulary	22	75
9. Use of Sources of Information	27	60
Total	6 hrs., 4 min.	716

standardized test results to teachers because Riverside, like many other school districts purchases a service for scoring statistical processing that takes about three months to return the results. Consequently, some teachers, especially those in the lower grades where the rate of development in reading is relatively rapid, find that when the test results finally arrive, they are not as instructionally relevant as they would have been had they gotten the scores earlier. The remedy, of course, is simple: have teachers or their aides score the tests at the school and then send the tests to the commercial center. Thus, teachers could get the results right away. Of course, this procedure involves duplication in scoring, increases the cost of testing, and may involve scoring and interpretation errors when done at the school, but the trade-off in having the test results become more useful may make it all worthwhile, especially if the school district could reduce the errors through inservice training of teachers and aides.

If schools were given the same test at the beginning and end of the school year, teachers could determine students' changes in achievement during the academic year. Although we can expect some student gain in achievement each year, the magnitude of change involves an interaction between the quality of instruction, the capabilities of students, and the components of basic skills. Only some of these factors are directly under teacher control. We shall discuss this further in the next section on the ideal use of test results.

Ideal Use of Tests on Basic Skills

Standardized Tests. The components of basic skills are two-fold. (1) acquisition of the skills, and (2) ability to use the skills. Acquisition of basic skills, such as learning to read, compute, or write, consists of objectives that approximately 95 percent of all students can theoretically learn if teachers provide adequate instruction and time for students to learn them.[31,32] For example, Figure 3 is an hypothetical illustration of changes in range of variability on a first grade test given to the same group of students in grade 1 and again in grades 2, 3, and 4. It shows that the students varied considerably in achievement at the end of first grade, but this variation decreased as most students had "mastered" the test by fourth grade, that is, approximately 80 to 95 percent had reached the 75th percentile or higher on the test. On the basis of such data we could infer that the students who reached this level of performance, had learned how to read at a basic level of proficiency and had shifted into emphasis on the use of reading to gain information or learn from text.

What is essential for separation of acquisition of a basic skill from its use is some consensus on what constitutes acquisition of a skill. Perhaps parents and professionals could reach the following consensus on a criterion for learning to read: Students have learned to read when they can achieve the 85th percentile rank in comprehension on a standardized, norm-referenced test for first grade. In other words, when they could, in effect, read and comprehend accurately all the paragraphs up to and including the paragraph, shown in Table 4, with its accompanying comprehension questions, taken

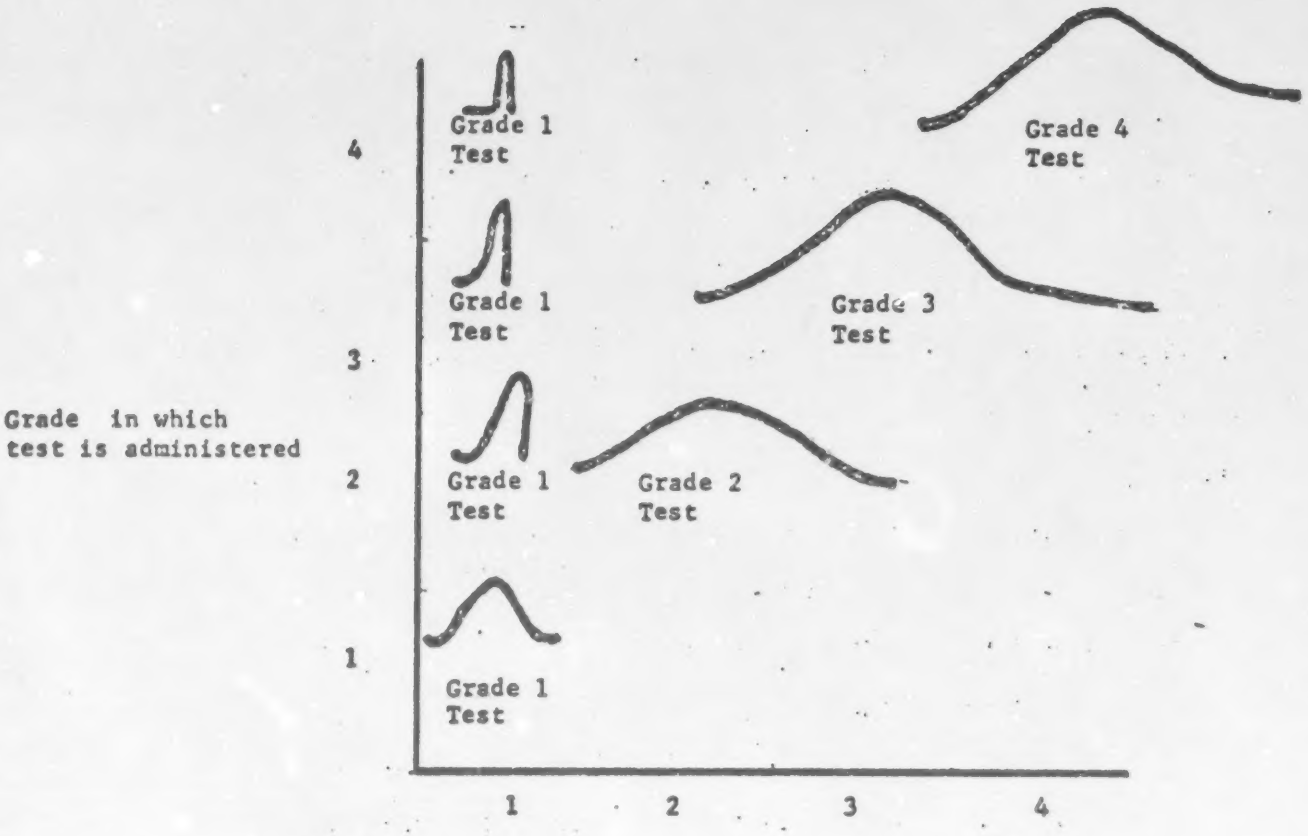


Figure 3. Repeated and grade level testing. The first column shows a norm-referenced first grade test or its equivalent given in grade 1 and repeatedly used in grades 2, 3, and 4 for assessing progress in learning to read. The diagonal shows a norm-referenced test for grades 2, 3, and 4, respectively, used at the ends of these grades to assess individual differences in use of reading for gaining information from written messages.

Table 4

Paragraph with Accompanying Comprehension Question from

The Metropolitan Reading Instructional Tests, Elementary, Form J [33]

On one afternoon the neighborhood children went to a magic show. They sat in the balcony. The crowd waited anxiously to see Mr. Moondrake. He stepped from the darkness with a roll of drums. He was a handsome, dark figure. A beautiful lady helped him perform the magic. For the final trick, the lady climbed into a large glass case. Mr. Moondrake covered the case with a sheet and fluttered his cape over it. The lady disappeared. Then he covered the case again, and the lady reappeared. It was scary. "What was the secret?" the children wondered. "Did the lady become invisible, or was she simply cleverly hidden?"

1. When did the children go to the show?
   - 0 At night
   - 0 In the afternoon
   - 0 In the morning

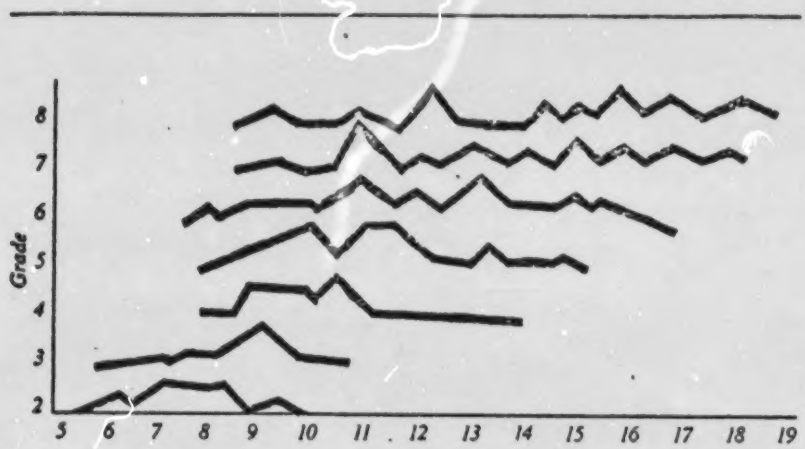
2. What happened to the lady in the case?
   - 0 She got lost.
   - 0 She disappeared.
   - 0 She was afraid.

3. Everyone wanted to know -
   - 0 who was invisible
   - 0 why Mr. Moondrake did the trick
   - 0 how the trick was done

4. The children were all very -
   - 0 excited
   - 0 angry
   - 0 unhappy

5. The best name for this story is -
   - 0 "The Neighborhood Children"
   - 0 "The Handsome Mr. Moondrake"
   - 0 "The Magic Show"

from the Metropolitan Reading Instructional Test, or a paragraph of equal difficulty level from any other standardized reading test. When students have attained this criterion, it does not mean that students have learned all they need to know about reading. They can still continue to improve, but their improvement is now more in the second phase of reading development, their relative reading achievement or ability to comprehend printed material. In other words, all basic skills have two components: an acquisition component and use of a basic skill for some purpose, such as gaining knowledge from textbooks, or solving a problem in mathematics, or creating a story, or writing a legal opinion. These applications of basic skills show increasing differences among a group of students as they progress through school, as shown in Figure 4 by the distribution of reading achievement scores on standardized tests for grades 2-8.[34]



Reprinted from E.F. Lindquist, ed., *Educational Measurement* (Washington, D.C.: American Council on Education, 1951), p. 13. Used by permission.

Fig. 4. Measured Range of Reading Comprehension Ages in Grades 2 through 8

Standardized achievement tests tend to subsume acquisition skills and increasingly assess use of the basic skill, such as reading to learn or gain knowledge from printed materials. But, instead of students becoming more alike over time as they do in acquisition of the basic skill, such as learning how to read, differences in use of a basic skill, such as reading to gain information from texts, should and does increase as students progress through the grades, as shown by the increases in reading achievement test distributions at successive grade levels.

Students, parents, teachers and the public should get two types of information: (1) progress of students towards acquisition of basic skills and (2) comparison of students with their peer group on standardized achievement tests. At the present time, schools only test and report students' relative achievement on norm-referenced tests. To also obtain acquisition type of information, it is necessary to administer the same first grade test repeatedly each year until students have attained an 85th percentile rank according to the test's first grade norms. This type of testing yields self-comparison information. The school could simultaneously give its usual grade level standardized tests. The results would then yield information on both phases of a basic skill. In reading, it would be: (1) learning to read and (2) relative ability to use reading to gain information from printed materials. We could plot this information for a hypothetical student, as shown in Figure 5.

We would interpret Figure 5 in the following way: The student was at the 20th percentile ranking at the end of grade one, but

improved steadily in reading acquisition, reaching the criterion level, an 85th percentile ranking, at the end of the third grade. We would then conclude from this self-comparison information that the student had attained the criterion level for learning how to read. However, in relative or peer-group comparison in reading achievement

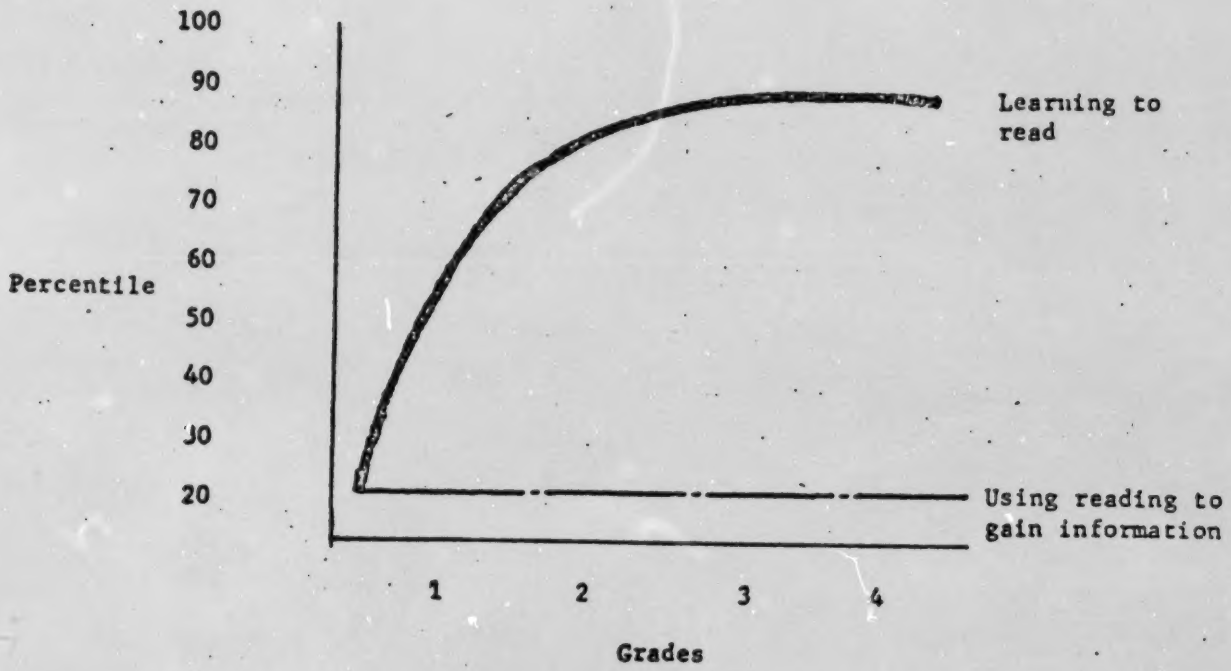


Fig. 5. Diagnostic information on an hypothetical student's progress in learning how to read and in use of reading to gain information. The "learning to read" curve is constructed by repeated administration of a first-grade, norm-referenced, standardized test. The "using reading to gain information" curve was based on grade level tests. Comparison of the two curves indicates that the student improved to a mastery level in learning how to read compared with a first grade criterion, but did not improve relative to his peer group in ability to use reading to gain information.

on tests standardized and norm-referenced for each grade level, the student's percentile rank remained at the 20th percentile rank at the end of each grade level. In other words, the student had learned how to read, but the student's reading achievement or ability to learn from printed material or texts continued to develop at the same rate relative to his or her peer group.

Although the student's relative reading achievement appears as a flat line in Figure 5, the flatness of the line does not mean that the student is not improving in general reading ability or ability to learn from texts. This misinterpretation stems from lack of knowledge of reading achievement tests and of percentile rankings. Reading tests for successive grade levels are so constructed that they become increasingly more difficult as successive grade levels. Their paragraphs and comprehension questions contain more complex sentences and more difficult words, assume more background information, and require more mature reasoning processes for making interpretations, inferences, and evaluations. Percentile ranks on norm-referenced tests show how well a student comprehends in relation to his or her peer group. To maintain the same percentile rank on succeeding tests of increasing difficulty, a student has to improve, but the improvement may be at the same rate as the peer group. If it is, then the student's percentile ranking will remain the same. That is, all the students may have improved, and so did our particular student, but his or her position in the group remained the same, just like a group of runners racing around a track can remain in the same order. However,

if a student's rate of development is greater than his or her peer group's, then the student's percentile rankings would increase and the bottom line in the figure would go up; if the student's rate of development was slower than his or her peer group's, the student's percentile ranking would decrease and the bottom line in the figure would go down.

Usually schools gather and provide information to parents, pupils, and professionals only on relative reading achievement. If this information is in the form of percentile rankings, the line in the figure frequently appears flat and results in the misinterpretation that the student is not making progress in learning how to read. However, if we gather and present information on both reading acquisition and relative reading achievement, we could prevent this misinterpretation.

We would also communicate more effectively in reporting test results to students and their parents if we include in the report the most difficult paragraph that a student could read and comprehend, that is, the highest level paragraph on which the student had answered 70 percent or more of the questions correctly. We can refer to this type of report as having content bound to the numerical information for each test administration; this type of report contrasts with content-free reports which provide only numerical information. For determining whether a student has improved in achievement as compared with a test given in a previous year, parents, pupils, and professionals could determine from comparison of this year's paragraph with last year's paragraph on a content-bound report whether the student had improved. This reporting procedure might increase the

cost, but it would make reporting of an indiviual's test results more meaningful.

When we experimented with this type of repeated testing and reporting of test results for a student at the second grade level, as shown in Figure 6, which depicts the results of self comparison and peer-group comparison, with and without content-bound information for a second grade student, educators overwhelmingly preferred this testing and reporting to current testing and reporting practices. Moreover, their positive attitudes toward the new testing and reporting practice was favorable, which contrasts significantly with their negative attitudes to current testing and reporting practices.[17,18] (For a detailed report on educators' preference and attitude toward tests and testing, see Testing Achievement in Basic Skills Volume I: Executive Summary and Research Reports.)

Schools also need to feed test results back into the curriculum and instruction. For this purpose a school district would need to have a plan to make this feedback meaningful and then employ or train a staff of consultants to teach principals and teachers how to use it, even if the tests did fit instruction very closely. The Riverside Instructional Support Services, which we have already described, is an example of this process. In short, modification of curricula and instruction are not likely to occur just by returning test results to districts and to schools. The CAP manuals for interpreting its test results provide suggestions for school faculties to consider in trying to relate test results to curriculum and instruction, but such suggestions are only a beginning to use of test results at the district or school levels.

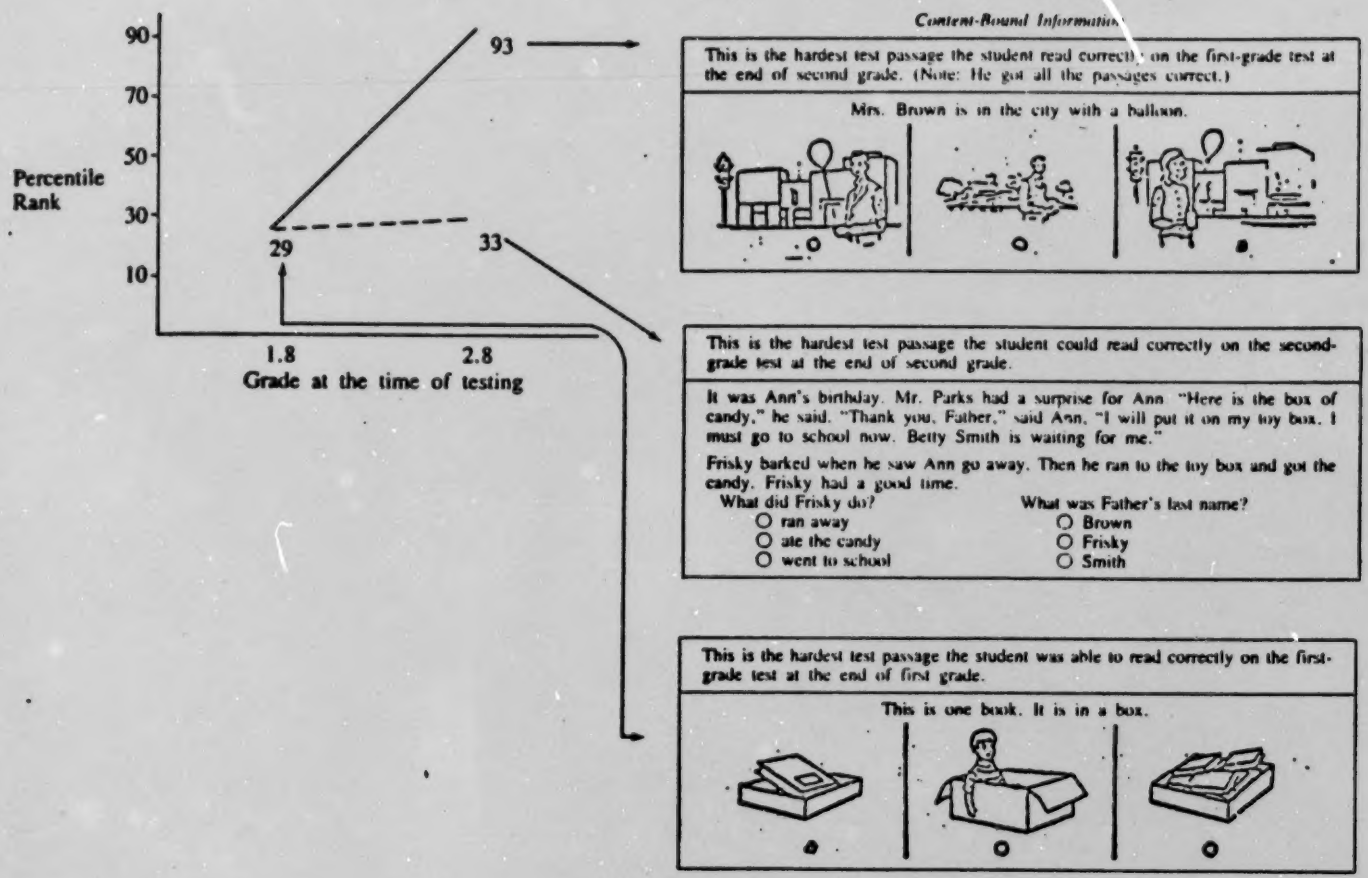

Content-Bound Information
90
70
50
30
10
Percentile Rank
93
29
33
1.8
2.8
Grade at the time of testing
This is the hardest test passage the student read correctly on the first-grade test at the end of second grade. (Note: He got all the passages correct.)
Mrs. Brown is in the city with a balloon.
This is the hardest test passage the student could read correctly on the second-grade test at the end of second grade.
It was Ann's birthday. Mr. Parks had a surprise for Ann. "Here is the box of candy," he said. "Thank you, Father," said Ann. "I will put it on my toy box. I must go to school now. Betty Smith is waiting for me."
Frisky barked when he saw Ann go away. Then he ran to the toy box and got the candy. Frisky had a good time.
What did Frisky do?
ran away
ate the candy
went to school
What was Father's last name?
Brown
Frisky
Smith
This is the hardest test passage the student was able to read correctly on the first-grade test at the end of first grade.
This is one book. It is in a box.

Fig. 6.[a] This graph shows the student's percentile rank on the reading comprehension subtest of the Comprehensive Test of the Basic Skills. At the end of first grade (1.8) this student was tested using the first-grade test. He scored at the 29th percentile, getting 12 out of 24 items correct. At the end of second grade (2.8) the student was given two tests. *First*, he was retested with the same test he had taken in first grade. This retest shows that when compared to students in first grade, he now scores at the 93rd percentile, getting 24 out of 24 items correct. *Second*, he was also tested with the second-grade test. He scored at the 33rd percentile, getting 11 out of 18 items correct. These data were presented in four ways: (a) only peer group comparisons, the dashed line only; (b) the self-comparison and peer group data, both lines as shown in the figure; (c) and (d) the same data as shown in (a) and (b), but with content-bound information, the most difficult paragraph the student could correctly read in each test.

[a]This figure is reproduced by permission from Singer and Dreher[17] and the National Reading Conference.

Pressures exerted by the community, judiciary, and local school boards on school districts to improve test scores on both CAP and district-adopted standardized tests have moved some elementary chools to expand instructional time devoted to "basic skills" at the cost of reduction or elimination of instruction in social studies, science, fine arts, and other courses that had been required prior to AB 356. For example, the San Diego Unified School District is under a court order to raise achievement test over a three years period (1980-1983) in its isolated schools (schools with more than 70 percent minority enrollment). In attempting to comply with this order, the school district has increased pupil "time on task" for instruction in basic skills, geared its instruction to basic skill objectives assessed on its standardized test (Comprehensive Tests of Basic Skills) and on CAP tests, and devised its own criterion-referenced tests for assessinng mastery of the objectives.

This narrowing of the curriculum is an inevitable result of the direction taken by the CAP testing program and the desire of school districts to perform at an expected or higher than expected level of achievement. Although the state testing program is only supposed to monitor, not drive and determine the curriculum, in practice it tries to do both. For example, the Examiner's Manual-1980 for the California Assessment Program's Survey of Basic Skills for Grade 3 states this "Caution Concerning the Use of These Tests;"

> The content of the Survey represents common skills taught in schools in California. Some skills tested may not be included in a particular

> schools' primary grade program; similary some skills taught in a particular program may not be tested. The inclusion or exclusion of a particular skill does not mean that the skills should be included or excluded from the curriculum at a particular grade; that decision must be made in the context of instructional objectives.

However, if a school district's scores on CAP tests are to increase, it should match curriculum objectives and instruction with CAP's instructional objectives. CAP encourages this match by reporting results for the grade 6 test in 143 skills areas (50 for math, 54 for reading, and 39 for written expression) plus total scores for each basic skill, all with expectancy bands. The report is highly specific; for example, the skill areas for capitalization lists separate scores for places, dates, and months! Not knowing which skill area influences total scores, school districts try to reach all skill areas.

CAP's suggestion to schools on how to validate CAP test results may lead teachers to match curriculum objectives to CAP test objectives as they answer this question posed in CAP's manual on Interpreting and Using Test Results (Survey of Basic Skills: Grade 3-1980, pp. 16-17):

> Are these skills [assessed on CAP tests] part of our curriculum?

Then the manual suggests that the school set this goal:

> Next year we will be within or above our comparison score bands in all three content area.

The only common sense strategy for schools to adopt, if they want to achieve this goal, is to make the school's curricular objectives congruent with the CAP tests' objectives. But if all schools were to try to adopt this strategy, it would be impossible for all of them to achieve the goal because they would all have to surpass their comparison bands. But, all schools cannot be above their comparison bands. To do so, they would all have to be above average!

Although the legislature may not intend to have the state control local district curriculum so tightly, the CAP tests which call for emphasis in basic skills are in effect doing so in a highly specific way, leading some schools to increase instructional time on basic skills, and consequently narrowing the curriculum. Is this consequence undesirable? The answer, at least for the intermediate grades is that it is undesirable. Why? Because comprehension in reading results from an interaction between text information and resources within the reader, including word recognition, vocabulary ability, general knowledge and reasoning processes. A narrow curriculum, which minimizes or excludes such courses as social studies, science, and fine arts is likely to limit students vocabulary abilities and general knowledge, and consequently students' reading comprehension abilities. Hence, we would predict that if this trend continues, we can expect comprehension to begin to fall on standardized reading comprehension tests as students progress through the grades.

What is the remedy? The CAP tests can expand to assess all areas of a broad curriculum, including social studies and science.

We have already proposed an increase in the scope of the CAP tests. In the process of interviewing the CAP staff, we suggested that the CAP tests should also assess social studies and science because of the importance of these content areas for reading comprehension. The CAP staff responded favorbably to this suggestion, obtained approval for it from the CAP Advisory Board on Reading, and revised the sixth grade test, and began to administer it in 1982. CAP now reports separate scorees for social studies, science, and literature in grade 6.

The expanded scope of the CAP test for the sixth grade is likely to prevent narrowing of the curriculum and perhaps will encourage some elementary schools to broaden their curricula. The consequence is likely to be a higher level of achievement in basic skills, particularly in reading comprehension. A similar change in the CAP test for 12th grade should also be made.

Proficiency or Minimal Competency Test

Another test which school districts must administer is the proficiency or minimal competency test now required by the Education Code as a prerequisite for the high school diploma. However, competency testing begins in the elementary school.

Legal basis for competency tests. The bill which began competency testing at the elementary school level (Grades K-8) was McAlister, AB 376 (January 31, 1977). The law governing minimal competency testing at the high school level had already been passed (Hart, AB 3408, Chapter 8561, Statutes of 1976). Both bills were superseded by further legislation (Greene, AB 65, January 31, 1977) requiring governing boards of elementary schools to articulate their minimal competency standards with the high schools' minimal competency standards. These bills and accompanying statutes ordered school districts to construct their own competency tests or else adopt a test and determine the score necessary for passing it by June, 1979. Each district has to assess pupil progress towards mastery of basic skills at least once during the fourth through sixth grades and at least once during the seventh through ninth grade and twice in grades 10 and 11. But whenever a student has demonstrated mastery of basic skills, the school does not have to reassess the student again. The new law in the Education Code directed the State Board of Education to provide a framework, including examples of basic skills standards, for assessing student proficiency, but specifically forbade the State Department of Education from constructing such a test (Education Code 51217).

The standards for ascertaining and evaluating individual achievement in basic skill areas were to include but were not limited to reading comprehension, writing, and computation in English necessary to success in school and life experiences. The boards were to

adopt these standards with the active involvement of parents who are broadly reflective of the socioeconomic composition of the district and the district's professional staff, including administrators, teachers, and counselors. The school also has to hold conferences with parents for any students who upon exit from grades 6 and 8 do not appear to be making satisfactory progress towards mastery of basic skills. At these conferences, the principal or principal's representative is to outline the instructional program for assisting the student in mastery of basic skills (Education Code 51216). Starting on June 1, 1980 no student can receive a high school diploma without having passed the school district's minimal competency test (Education Code 51217) and completing the school district's course of study, which has to include English, American history and government, mathematics, science, physical education (unless exempted), and other courses as may be prescribed. However, the minimal competency law stipulated that school districts were to adopt differential standards for students with diagnosed learning disabilities; the Educational Code already permits alternate means for these students to complete the prescribed courses of study including practical demonstration of skills and competencies, work experience, interdisciplinary study, independent study, and credit courses at postsecondary institutions.

Contribution or Adoption of Minimal Competency Tests. To satisfy the legal requirements, some school districts constructed their own tests. The Los Angeles Unified School District estimates it spent some three million dollars to develop its own test and create miniunits for instruction in basic skills. The district then

contracted with the McGraw-Hill Book Company to publish the test and miniunits. The reading test is known as the Senior High Assessment of Reading Proficiency (SHARP). Some school districts adopted a competency test, such as the Educational Test Service's (ETS), Basic Skills Assessment. Table 5 contains a description of each of these tests. But most districts constructed their own competency test.[13]

Evaluation of Two Minimal Competency Tests

The minimal competency tests are criterion-referenced. According to ETS's description of its Basic Skills Assessment test, it is "relatively easy for the population as a whole." Using the Flesch Reading Ease formula to determine the reading difficulty of a sample passage in ETS's description of its test, we found that the passage has a readability of grade 8 to 9, which means the passage is about as difficult to read as a passage in Reader's Digest or Time magazine. About the closest we could come to a 100 word passage on the SHARP test was (1) a notice for a rent increase which was "fairly difficult" to read, is equivalent to reading material at grades 10-12, and is comparable to the reading material in Harper's and Atlantic magazines, and (2) a driver's license manual which contained reading material that was "fairly easy," is equivalent to reading materials at grade 7, which according to the Flesch norms indicates the reading material is about as difficult to read as a slick magazine such as True Stories. On the math section of the ETS test, the item with the highest level of difficulty is computation of percent, which schools typically teach in 7th grade math classes. The items in the writing part of the test

Table 5

Descriptions of Two Proficiency Tests

A. Senior High Assessment of Reading Proficiency (SHARP). Monterey California: CTB (McGraw-Hill, 1975). Copyright© 1975 by Los Angeles Unified School District.

Purpose: Objectively assess reading skills for forms and documents adults read and/or complete continuously. Total number ofitems: 120. Passing score: 81 (67 percent of the total).

Categories:

1. Can you follow directions for voting, driver's tests, library cards, job and rental applications, unemployment insurance claims, letter writing, social security documents, and change of address forms. Example: Rent receipt followed by questions with multiple choice answers.

2. What's In It? Reading and comprehension of signs, labels, and want ads, Yellow Pages, dictionaries, newspapers, warranties, bank accounts, signature cards, credit applications, and check writing. Example: Driver's License Test Signals: Section represented from state's vehicle code followed by questions and multiple choice answers.

3. What Does It Mean? Reading skills involved in the use of maps, area codes, cash register tapes, bank statements, TV guides, income tax forms, utility bill codes, and job resumes. Example: Memo on rent increase and electricity payment policy change followed by questions with multiple choice answers.

B. Basic Skills Assessment. (A cooperative effort by Educational Testing Service and a national consortium of school districts, 1977).

Description: Three multiple choice tests, each takes approximately 45 minutes. The Writing Sample, a direct measure of writing ability, is read and scored by trained staff with backgrounds in English.

Purpose: Identify between group requiring remediation in basic and group not requiring such remediation. Test reliabilities are over .90. Assess school skills and life skills.

1. Reading (Total items: 65)

Table 5 (Cont.)

Skills and items: Literal Comprehension (33), Inference (25), Evaluation (7).

Content: Consumer (12 items): telephone directory, loan agreement, guarantee, advertising, operating guides, product information.

Learner (14 items): Newspaper or magazine, narrative, dictionary, school catalog, cartoon, book or periodical titles.

Citizen (14 items): Editorial, tax form, driver's application form, law, political propaganda, community resources.

Protector (12 items): Medicine label directions, product warnings, nutritional information, first aid information, road map.

Producer (12 items): job application, description of benefits, bus schedule, want ads, work-related information.

2. A Writer's Skills (Total items: 75)

Skills and items: Mechanics of writing (spelling of common words, 14; capitalization and punctuation, 14); effectiveness of expression (appropriate English usage, 16; such as, subject verb agreement, verb form, tense, etc.); Evaluation and organization (27 items such as clarification of ambiguities, relevancy, sentence relationships, economy, logical sequence of information, diction according to tone).

Content

Consumer: Letter explaining a problem, personal check.

Learner: Informative passage, narrative selections, descriptive passage.

Citizen: Announcements and messages, letter expressing point of view, properly addressed envelope.

Protector: Letter requesting information, announcements and messages, instruction.

Producer: Want ads; letter of application for employment, employment form, business correspondence.

Table 5 (Cont.)

3. Mathematics (Total items: 70)

   Skills: Addition, subtraction, multiplication, division of whole numbers, and fractions; finding percent, compute average, estimate and compute measurement, read graphs, compute interest, compute elapsed time.

   Content: Computation (30 items), Applications (40 items) Consumer buying, taxes, banking, wages and salaries, household expenses, shipping and mailing, transportation, statistical inference, health and nutrition, sports.

4. The Writing Sample. (3 types of writing; suggested time: 45 minutes). Specifications for writing exercises:

   1. Write a letter applying for a job in response to a want ad.

   2. Fill out a simple form, such as an application for a driver's licence.

   3. Choose one:

      a. Creative writing: Write imaginately about a written or visual stimulus.

      b. Informative writing: Accurately convey given information, such as relaying a telephone message or describing a process.

Scoring:

1. Analytical scoring (points) of some exercises by teachers aides

2. Holistic scoring (judgment or total impression; ignore minor grammatical errors) of some exercise by English teachers

also do not appear to be any higher than content taught in grades 7 to 9. Thus, these two tests are at a difficulty level appropriate for an average student in grades 8 to 9.[35]

However, in California each school district determines what the passing score on its competency tests will be. The percent who do not pass at the end of grade 12 varies from district to district. Hence, in California, the meaning of minimal competency varies from school district to school district. Data are not yet available on how many 12th graders in California failed to receive a diploma in June 1981, but the estimate is about 7 percent, some 21,000 students.[13]

National context. Some 38 states now have minimal competency testing. They vary in government level for setting of standards (7 local, 19 state, and 12 state and local); grade levels assessed (grades 1 to 12); expected use of standards (grade promotions, 3; high school graduation, 16; and local option, 2); early exit (only 2, California and Florida), remediation (11, including California), and first graduating class assessed (1980 to 1985).[36] To make these figures more meaningful, we shall contrast New York with California. New York adopted one test for the entire state and set the passing score for it. It assesses students in grades 3, 6, and 8-12 and uses the results for determining who needs remediation and for high school graduation, starting in 1979.

In contrast to California's local school district adoption of tests and selection of passing scores, all high school students in New York take the same test and have to pass it with the same minimum score. Hence, New York's minimal competency testing program was more

economical to set up and is uniform in standards for passing the test.

A consequence of the differences in governmental levels for determining minimal competency standards for high school graduation is that parents in New York perceive that the governor and state legislators are the people who determine whether their children will receive a diploma upon completion of the 12th grade; but in California, it is the superintendent and members of the local school board who make this decision and have to bear the brunt of any negative parental reaction.

Competency tests did not suddenly arise. They have a background in the past 30 years of American history.

Historical Background. The imposition of a minimal competency test is the result of a social, economic and political shift in attitudes in the country in the 1970s that resulted in "holding teachers accountable," particularly for development of basic skills. It also is a reaction to the 100 percent promotion policy, also known as as social promotion, initiated in the 1950s. Studies done in the 1940s and earlier had shown that students who were not promoted did not achieve as well as comparable students who did progress with their peer group.[37] In the interests of achievement, as well as social adjustment, schools decreased their retention rate. This decrease also coincided with an influx of students in the schools, the post-World War II baby boom, which formed a demographic bulge that progressed through the schools and led to unprecedented expansion at each school level, elementary through university, for the

next 20 years! Contraction of the school systems at these levels and termination of a teacher shortage became a phenomenon in the 1970s and appears to be continuing through the 1980s. Expansion is psychologically and sociologically, but not economically easier for society; contraction tends to have the opposite effects. But contraction did make space available in schools for retention to again become feasible. However, retention is more costly since it tends to increase the time that retained students are in school, unless they begin to drop out again at a faster rate.

The competency test is also linked to the curriculum and to an increase in emphasis on basic skills, particulary their applications to social competence, known as life-skills or functional abilities. However, the imposition of competency tests has social and legal consequences.

Social and Legal Consequences

Those students who do not pass the minimal competency test now comprise one tier of students who have gone through school. Another tier consists of those students who take and obtain scores on college placement tests high enough to make them admissible to four-year colleges and universities. In-between these two tiers of students are a third group. In California, the two lower tiers may go on to two-year community colleges if they want to continue in public education.

Although minimal competency tests are now in widespread use throughout the country, they have been and are susceptible to legal challenges and to criticism as to the wisdom of using such a test.

For example, M. S. McClung, director of the Center for Law and Evaluation, Harvard University, has questioned their legality, particularly the tests' lack of instructional validity and other due process issues.[10]

Furthermore, the International Reading Association has taken the position that a single reading test should not be the sole criterion for evaluating minimal competency in this basic skill after 12 years of education.[38]

Minimal Competency Assessment in California

The California State Department of Education's 1980 status report to the Legislature on implementation of California's Pupil Proficiency Law found that (1) 78 percent of school districts constructed their own proficiency tests, (2) 47 percent assessed both "school skills" and "life skills," 39 percent emphasized "school skills," and 15 percent stressed "life skills," (3) 90 percent of the school districts indicated greater emphasis on instruction in basic skills, (4) few districts reported inservice training of teachers for instruction in basic skills, (5) less than half the parents of failing students attended conferences the law requires schools to hold, (6) 96 percent of 11th grade students knew they had to take proficiency tests, and 75 percent felt they were about the right level of difficulty while only about 12.5 percent thought they were too easy and the other 12.5 percent thought they were too hard, (7) only 58 percent of the school districts had adopted a differential standard for students in special education programs, (8) an estimated 7 to 8 percent of twelfth graders (about 21,000 students) are expected to fail the proficiency tests and

not receive diplomas at graduation in 1981, which will put a strain upon superintendents and school boards to lower the requirements for passing or else make the tests easier, (9) the drop-out percentage may increase as a result of an interaction between the school's course or unit requirements for graduation and the failure to pass the proficiency test, (10) 30 to 50 percent of the sophomores (100,000 to 150,000 students) were failing one or more sections of the proficiency tests and were enrolling in one or more remedial classes which is endangering and perhaps "trivializing" the high school curriculum as math and English teachers have to teach more remedial courses and hence have to reduce availability of advanced courses (italics not in the report).[13]

The last item indicates that a large amount of school resources has been shifted from the top 70 to the lower 30 percent of the students in order to comply with the requirements of AB 65. This shift in school resources represents a drastic change in the quality of the high school's program which is already less demanding than in other states.[1] Although offering more remedial courses may help some students at the lower end of the achievement distribution pass the proficiency test, it is likely to be counterproductive to improvement in basic skills on the CAP tests for the upper 70 percent of the students. Their improvement depends on development in advanced courses in the various content areas of the curriculum. As teachers have to administer proficiency tests, participate in planning

conferences for students who fail the tests, and offer more remedial courses in English and math, the offering of advanced and elective courses is likely to decrease with detrimental effects to the quality of high school education and preparation of students for the colleges and universities.

Although high schools are also charged by the Educational Code to prepare students for colleges and universities, achievement in this program is not promoted by the State of California, but it should be. If the trend towards lowered achievement in CAP test scores at the 12th grade level and reallocation of resources to the students who fail the test continues, we are also likely to witness an exodus of students whose families can afford private schools. A recent report, which indicates that achievement in private and parochial schools exceeds public schools, is likely to accelerate this exodus even though the initial report was essentially biased and misleading because the college-bound students in private and public schools were actually equal in achievement.[39] But, if parents believe that private education will better prepare their children for college, and particularly if tax credits for private education or a voucher plan that gives parents a choice of the use of a voucher for paying for private education is legislated, then we are likely to witness the creation of a two-class system of high schools, such as Australia now has, with public schools primarily for students from poor families. Development of such a system would be detrimental to the ideals of education in a democracy.

Although schools have to make decisions based on proficiency tests, we should recognize their limitations. Even though they are criterion-referenced and closer to classroom instruction, they assess the effects of instruction offered initially at elementary or junior high schools. They also have other limitations for decision-making, as do norm-referenced tests.

Gap in Monitoring State Programs

The University of California (UC) and California State University and Colleges (CSUC) require applicants to present Scholastic Aptitude Test scores, a high school diploma with certain required subjects, and at least a minimal combination of SAT scores and grade point average. The University of California also requires three College Entrance Examination Board achievement test scores. CSUC tests its students on an English Placement Test. Otherwise, neither UC nor CSUC monitor the quality of high school preparation. They both accept at face value the courses and grades listed on high school transcripts. No effort is made to determine whether a particular high school is lenient or stringent in grading its students. Hence, neither UC nor CSUC knows whether entering students have, in fact, attained entry level knowledge that first year courses presume students command, except in English and for UC in two additional subject areas.

What we do know is that California students in 1978-79 "studied academic course work for fewer years while in high school, were enrolled in smaller percentages of honors courses, yet were given higher grades and have higher expectations for advanced college placement" than the average student nationally who took the College

Entrance Examination Board's test.[1] The legislature and the public needs to know how this relative redirection in academic preparation is affecting the achievement level of students preparing for colleges and universities. Although the Educational Code (51224) requires schools to prepare students for admission to higher education, the state does not monitor the program. It only monitors two of the three student tiers: the low achieving student in the profiency program and the average student's level of achievement in basic skills. Thus a gap exists in state monitoring of mandated programs.

Limitations of Tests for Evaluation and Decision-Making

We can use tests to decide whether students have reached a criterion, infer whether students are making rapid or slow progress, or even determine their status in relation to their peer group. The test results may even reflect upon current instructional programs. But we can not use test results alone to determine the merit of a program. We must base such a decision upon a whole host of steps, which includes needs assessment, program planning, implementation, evaluation, progress evaluation, outcome evaluation, and frequently a research study.[40] For example, CAP indicates the median sixth grade student ranked at the 55th percentile, but the median 12th grade student ranked only at the 34th percentile in reading on national norms. We would have to formulate and test various hypotheses to determine why the drop-off at 12th grade occurs. In other words, test scores only answer the question of what is the current status of achievement, they do not determine cause-and-effect relationships nor do they determine a solution for improving achievement. Research of

the type that the California State Department of Education conducted in its California School Effectiveness Study is necessary to answer a causal type of question.

Furthermore, we need new tests, particularly those we can use for diagnosis of knowledge and processes in reading, writing and arithmetic achievement. Some computer programs are already available for such diagnosis. These programs not only determine computational errors in arithmetic, but also provide reasons why students make some consistent errors. This kind of information is very relevant for a a close fit between teaching, learning, and testing.[20]

We shall go into greater detail in our next section on criteria for constructing tests that will fit in more closely with instruction and learning at the classroom level. As we have pointed out earlier, when we get closer to the classroom level, we need tests which provide teachers with information that is appropriate for the instruction of individual students.

Classroom Testing

Purpose of Testing

Standardized tests, such as CAP and commercially-published norm-referenced tests sometimes provide information that is useful for evaluating, understanding, and improving learning. But, as we explained earlier, standardized norm-referenced tests are most appropriate for answering questions on comparisons of students, school districts, and states. However, some standardized tests, such as the Metropolitan Instructional Reading Tests, have been careful to relate objectives and items to school curricula and to report information more useful to teachers.

Nevertheless, norm-referenced tests have their limitations. They frequently reflect only comparisons on relative achievement. Such comparisons can be misleading. For example, some students may actually have increased in absolute achievement (they could read and comprehend more difficult passages or compute and solve more complex arithmetic problems), but decreased in percentile rankings because the norm group had increased in achievement at a faster rate. Moreover, when standardized tests report subtest scores or items by objectives scores, the subtests are too short or the test items are too few for the scores to be reliable. Also, the reliability given in the test's manual is frequently only for the total test on the total sample. Yet we know that the reliabilities on the subtests are lower and the reliabilities for different subgroups, such as high and low achievers, can vary within the same test considerably. However, newer ways of

constructing and interpreting tests, such as CAP first started to do in 1980, tends to ameliorate some of these reliability and interpretation problems. However, teachers have difficulties in interpreting test results. They apparently have a 100 point scale in mind, regardless of how test data are reported.[22]

Essentially norm-referenced tests indicate only current status. Their results may "raise a red flag" to signal the teacher to undertake further investigation to determine why this student or this class obtained this status. Commercially-published tests also leave unanswered many other questions, especially what, how much, and under what conditions students learn or fail to learn. Moreover, they do not differentiate an error due to a student's lack of background and an error due to a deficiency in a particular skill. Nor do these tests help to determine what changes teachers need to make in teaching, or students in their motivation and strategies for learning, or schools in their course requirements and content. Clearly, at the classroom level we need different types of information to supplement or replace the data we get from commercially-published, norm-referenced tests.

Classroom testing also has other purposes to satisfy. Chief among them are placement, diagnosis, and mastery. Placement at the classroom level refers to the decisions a teacher has to make. For example, teachers have to answer questions, such as these: "In what level of graded textbook should I place this student?" Or, "what part

of a skills continuum is most appropriate for this child?" Diagnosis is the identifying of a particular difficulty a student is having in learning a particular skill. It is basic to selecting specific learning experiences for the student. Mastery is determining whether a student has learned a specified amount from a learning experience. The amount of knowledge or skill a student is to acquire as an indicator of mastery at one stage is generally based on what the teacher or the curriculum requires a student know and be able to do at a subsequent stage or be able to apply in life-like situations.

Placement tests only determine a students' initial instructional level. Students placed at the same initial level may and usually do progress at different rates of learning. Consequently it is necessary to give mastery tests to indicate whether or not satisfactory achievement followed instruction and differential rates of progress. If satisfactory progress does not occur, then the teacher should give a diagnostic test to determine the kinds of difficulty and the reasons why. In remedying diagnosed learning difficulties, teachers have to use different approaches for different difficulties.

Uses and Limitations of Tests and Test Results

Teachers realize tests and testing are necessary for quality education, but the conflicting demands for classroom testing that authorities at the local, state, and federal level have imposed upon teachers are making them confused and annoyed. Many of them perceive externally imposed tests as irrelevant to the purposes of classroom testing and to instruction. In a survey of teachers in one large school district, Edelman found that about 20 percent reported

principals had asked teachers to change their teaching to better prepare students for mandated tests, but the tests results did not influence teaching because many teachers reported they did not see them. When they do see them, most teachers reported the tests were less accurate than teacher evaluations as indicators of pupil abilities. Sixty percent indicated that the Comprehensive Test of Basic Skills (CTBS) should be revised or replaced; 48 percent felt that way about the CAP tests and 24 percent about the Survey of Essential Skills (SES), Los Angeles Unified School District's own criterion-referenced test for assessing minimal competencies. Only 39 percent wanted to continue using the SES, 20 percent wanted to keep the CTBS tests, and only 7 percent voted for the CAP tests. Perhaps the vote on the CAP tests is so low because teachers know that CAP tests will give them no information on individual students. Unlike the other tests, CAP does not have any student take a complete test. Over 90 percent thought the same information could be obtained by using fewer tests and that the benefits from the testing program were not equal to the costs in teacher/pupil time and money. The teachers felt this way about the tests even though 62 to 74 percent of them had attended meetings to acquaint them with CAP, CTBS, and SES.[41]

Attitudes of teachers towards tests and testing are important because they not only prepare students for tests, but also administer the tests. The pressures that operate on teachers or that they feel or perceive may also affect how they give the tests and how students take and perform on the tests. If they do not motivate students to do their best on the test, a basic assumption underlying any achievement

test score will not have been met. Or, if teachers depart from test conditions by promising a reward to a class for doing well on a test, the interpretation of the results is not likely to be completely valid. In general, if students do not take a standardized test under the same conditions under which it was standardized and normed, it would not be appropriate to use the test's norms for interpreting these students' test scores. The reliability and validity of test results are also affected by the kind and degree of coaching in how to take tests; some teachers may not separate their roles as teachers from their roles as testers and confuse the two roles by coaching students taking the test. Teachers also have to exercise judgment on whom to exclude or include from taking a CAP test. If they are stringent in applying exclusionary criteria to little or non-English speaking students so that fewer students are excluded, the school's expectancy band is likely to be lower; if they are lenient, it is likely to be higher. The result is a variation in expectancy bands and achievement levels of schools and hence in determination of whether a school is achieving below, at, or above its expectancy level. Teachers may also vary in their definitions and interpretations of how to prepare children for a test without teaching for a test. Even a relatively minor departure from a test condition such as asking all students to complete each question, even if it means guessing, shows an apparent gain in test scores. Thus variations in teacher attitudes, judgments, and interpretations, and

in the conditions under which tests are given can affect the reliability and validity of test scores.

Some teachers perceive that some CAP tests and some standardized tests do not appear to measure, at least on the surface, what the test purports to measure. That is, the content of the test does not correspond to what they consider are basic skills. For example, in the CAP test for measuring written expression, pupils do not actually write, but instead they recognize sentences, identify errors in punctuation, and select from among multiple choice answers the word that best completes the sentence. In short, the test does not have "face validity."

Further, many teachers are frustrated because CAP lacks a clear description of the items that assess the tests' objectives and the percent of students at each grade level who pass the items. Even if

teachers had this information, they would still find it difficult to improve instruction because they do not know which students need to improve on which objectives. However, if teachers had item-objective information, they could construct their own criterion-referenced tests to determine the needs of their students and after instruction assess whether they have achieved mastery. Without the item-objective information, the teachers cannot even begin to make tests, such as the CAP tests, instructionally relevant.

Minimal competency tests are also having unanticipated consequences. Just as some teachers know how to increase their class's average scores on a standardized test by concentrating on high achieving students, because the high achievers tend to gain the most from instruction and consequently raise the class's average score on the test, they have now learned where to concentrate their instructional time for the class's improvement on competency tests. They can ignore high achievers once the high achievers display competency in the minimal skill and instead give major attention to helping low achievers pass the minimal competency test.

Even though some teachers have become test-wise, they are still annoyed at the multitude of tests they must administer. In addition to CAP tests, federal and locally-required standardized tests, mandated competency tests, and their own tests for assessing classroom instruction, they also must administer criterion-referenced tests which accompany particular text materials. Unfortunately many of the tests accompanying the text materials are not acceptable measuring

instruments. They are unreliable and contain few items per skill objective; the items are very specific to the text and therefore do not measure generalizability, the ability of the student to apply knowledge or a skill in a variety of situations. Fortunately, it is possible for publishers and test-makers to correct these shortcomings of tests and produce tests that are closer to the ideal for classroom use.

Ideal Types and Uses of Tests

Classroom teachers need tests yielding data with which to plan instruction and select instructional strategies. That is, testing should not only provide information on what students know, but also on what methods are appropriate for their further development. For example, we know that individuals need a pattern of at least three subsystems (sounds for letters and words, semantics, and syntax) for attaining reading comprehension. If testing reveals that two of these subsystems are better developed, then the teacher can select a method that will improve the weaker subsystem. We also know that teaching elementary students summarization strategies will improve their comprehension on CAP tests; this knowledge indicates teachers need more appropriate instructional strategies and diagnostic tests to determine whether students have learned and are using them.[30] We need to develop more tests that are useful for such diagnosis and instructional strategy selection. Therefore, we propose criteria that all tests should satisfy if they are to be ideal

for classroom use. These criteria are not only appropriate for CAP tests but for all standardized norm-referenced tests.

Statewide Tests

Ideally, standardized norm referenced tests, which include the CAP tests along with commercially-published tests, should possess seven criteria of importance to classroom use. We shall state these criteria and use them to evaluate the CAP tests.

1. Descriptive rigor. Standardized tests should offer an unambiguous description of what the tests are measuring. Different teachers should be able to consult a test manual and draw identical conclusions regarding what the test items are assessing. The 1980 CAP manual seems to meet this criterion by providing definitions and examples of items.
2. Validity. Standardized tests should display several types of validity:
   (a) Domain selection validity. Information on the process for selecting skills measured by the test should be stated in the manual. Specifically, the manual should indicate who selected the competencies and on what basis. CAP provides a very general description for meeting this criterion.
   (b) Descriptive validity. The test manual should contain a summary of the degree to which test items are congruent with the test manual's description of these items. CAP provides only sample items.

(c) Item validity. The test manual should contain data on how well pupils at various grade levels perform on the items in the test and on the objectives assessed by them. The CAP tests only provide general, not specific information on this criterion.

(d) Curricular and instructional validity. Does the test assess objectives in the curriculum and taught in the classroom? The test manual should provide adequate information for teachers to judge whether the test satisfies these criteria. CAP would better satisfy this criterion if it would provide a listing of items grouped by objectives.

3. Reliability. (a) If teachers are to make decisions for their programs on the basis of test results, the test manual should provide some evidence that the test yields data which results in decision consistency, that is, whether the same teachers arrive at the same instructional decisions when given the same data at two different times. (b) If the test is for use with individual students, the teacher needs to know that the test will consistently identify students who have or have not mastered the skills measured by the test. The CAP tests are not for individuals but for school and school district assessment. Hence, they can only show decision consistency at these levels. Although CAP claims it is useful at these levels, it has not conducted any study on decision

consistency.

4. Comparative data. To judge how well an examinee should perform with respect to a skill in question, teachers need comparative data regarding how other students have performed on the test, not only on the total test, but also on its subtests and on its items grouped by objectives. CAP does return data to school districts and to schools with breakdown of information in numerous ways which do allow for comparisons within and between schools, tests, and group characteristics.

5. Norms. The test's manual should have a complete description of the sample of students who participated in the norms for the test. This information should satisfy criteria for determining that they are a representative sample. CAP test do not sample, but exhaustively test all students in grades 1, 3, 6, and 12 in California. However, when they are interpreted according to national norms of standardized tests then the representativeness and the timeliness of the norms on these other tests should be examined.

6. Clarity of directions. The manual for administering the test contains clear directions for the teacher to give to the student. The printed instructions should have examples for students to do, and the test should have examples to further clarify the directions. Whenever possible, the directions should allow the teacher to demonstrate how to take the test.

The directions for the CAP test appear to be quite clear.

7. Motivation. The purpose for taking the test and use of the test results should be explained to students and their parents. Both of them should get the results and an interpretation of them. The CAP tests fail on the latter criterion because individual test results are not returned to students, their parents, nor their teachers. The results are not returned because no student beyond the first grade takes a complete test.

School District and Classroom Tests

We propose the following purposes for construction or selection of tests for use at the school district or classroom level:

1. School districts and teachers sometimes need to compare or rank students with each other. School districts or teachers sometimes need to select some students from among many, such as selection of students for entrance into special programs, colleges, graduate schools, and graduate programs. In these situations, tests which by design intentionally discriminate among students according to ability or achievement play an important and necessary role in admissions. For these decisions, standardized tests in current use are appropriate
2. Evaluation and improvement of classroom instruction. We need tests that reflect actual instruction and assess what students have learned in the classroom. The most sensitive tests for this purpose are reliable and valid criterion-

referenced tests.

3. Information about individual learners. We need information about individual learners in order to determine instruction that is appropriate for them. For this purpose, teachers need to use such devices as informal reading inventories, diagnostic tests, and criterion-referenced tests geared to their own instructional materials and objectives.
4. Decisions about individuals require reliable, valid, and analytical information about individual learners. Reliability of some tests may be high enough for research or group placement, but not for making recommendations about individuals. For example, a test's reliability can range from 0 to 1.0. A reliability of .80 is high enough for research or for making decisions about groups, but not for making decisions about individuals; for this purpose, a test should have a reliability of .90 or higher. Very few tests have a reliability this high. Two which do are the Stanford-Binet Test of Intelligence and the Wechsler-Bellevue Intelligence Scales when administered by competent and conscientious testers to motivated, cooperative, attentive, and emotionall stable students. Also, the validity of a test's items and objectives and their relation to instruction weighs heavily in the value of an achievement

test for making decisions about individual learners and their future instruction. Although teachers may construct criterion-referenced tests for assessing student progress, we also need studies to determine the relationship between test information and instructional strategies. For example, should a teacher use a strategy for associative learning or one for conceptual learning, or another for rule-governed or deductive learning, or all three in a sequence? When or under what conditions does a teacher use one strategy or another?

5. Tests useful at the school and classroom level should relate directly to the school's or the teacher's objectives, and not vice-versa. We should test learners on what the Education Code and the local governing board want students to study. Nationally distributed standardized tests have some difficulty in meeting this criterion. When education is locally controlled, it tends to have some aspects in common with other school districts in the state because of the impact of the Education Code but it also has some diversity which reflects the local community's values, interest, and needs.

   For example, we would expect to find some differences in curriculum between rural and urban schools. But when parents, administrators, and legislators put pressure upon teachers to obtain higher test scores, the teachers tend to

respond by narrowing the curriculum or even adapting or modifying the curriculum to fit the test's objectives. In this event, the test determines the curriculum.

6. At the same time, tests should not always be specific to the teacher's or the local school's objectives. At the classroom, school, and district level, we need to know whether the local curriculum is the most effective for the educational development of the students. By use of nationally standardized, norm-referenced tests which claim to assess a curriculum common to the country and have norms on a representative sample of students, it is possible to make inferences about the effectiveness of a local curriculum and instruction, if the students are comparable to the national sample. It is also possible to determine from test data that some curricular objectives are missing from or play a weak role in the curriculum. For example, some competency tests have demonstrated that development of life-skills, such as ability to read and apply driving instructions, fill out application forms, and compute bills need more emphasis in schools for some students. Minimal competency testing has focused attention upon these educational objectives.
7. Test data should provide information about the subskills of educational achievement. Survey tests usually communicate only comparative information. Diagnostic tests can determine

an individual's level of development in subskills underlying basic skill achievement. Teachers need to use both types of tests because the information they get from them is complementary. The CAP tests are now trying to supply both types of data, but these data are only useful primarily for assessment of the district and the school's curriculum. The district needs to give its own tests to get information for individual students' current status and subskill development. Some group tests, such as the Metropolitan Reading Instructional Tests, are appropriate for this purpose.

8. Test data should be communicated in ways that are understandable to consumers. Often those who administer and report test data do so in a content-free way. They report data in the form of stanines, percentiles, averages, etc., without providing consumers with the content or items to which students responded that resulted in the data. Consumers need to have content-bound data so that they will understand what students know and are able to do along with how much they have progressed towards certain objectives. Even when content is present, the numbers associated with it might not be understood. For example, unsophisticated test consumers may understand the meaning of percent correct, but do not comprehend as well the meaning of percentiles. They should know that half the students will have percentile rankings above 50 and half will be below 50 because

percentile ranking always results in this distribution of scores. The same result obtained for school districts when CAP renormed tests scores each year. Even though consumers may understand averages or percentile rankings, they still want students, classes, schools, districts, and the state to be above average, above the 50th percentile. Apparently nobody wants to be average or below average. But 50 percent always will be. The only way to get higher rankings for more than 50 percent is to keep the test norms constant; then, if a target group of students improves, more than half could get rankings higher than the 50th percentile. For example, if the state constructed percentile rankings on the 1980 test and kept these norms, and if achievement in the state went up in successive years, the average achievement within the state would be above the 50th percentile ranking.

An analogy can be made to a group of people running a footrace around a track. So long as we rank them in order, one will be first, one second, and so on. Good coaching and diligent practice can increase their speed and can cause some of them to pass others, but in the end, only one can be first, another second, and so on. Someone must be last. Measuring who is first, who is last, and who is in the middle is a poor way to decide if we have good coaches or fast runners, much less whether we have invested our funds wisely in coaches, equipment, practice, food, and medical care.

Rather than fire the coaches or question the heredity of the runners, we should ask whether we are testing their performance in a way that makes it possible to determine improvement. The answer is no; we are not. If we could use an absolute measure, such as time measured with a stop watch, we could measure performance in terms that will show whether each learner improved or declined regardless of the achievement of other learners. CAP claims that its tests, beginning in 1980, will yield such information, but if its claim is valid, it still only applies to school districts, schools, and grades, not to individual students. Even so, the CAP tests will not be able to determine what level of achievement is possible; interpretations of CAP test results will still have to use comparative information to determine how California achievement ranks against other states' achievement.

9. Test data should help school staff evaluate intervention programs. The relationships among the program, students' abilities, and achievement outcomes comprise the essence of evaluation. Although test information can provide data on student achievement, it will not determine cause-and-effect relationships. We still have to discover ways to relate test results to instruction before we can succeed in using test results for making instructional decisions that are reliable and valid.

10. Teaching, texts, and tests should be coherent. A well orchestrated program in development and assessment of basic skills should be founded on a clear definition of the basic skills, instructional materials that are valid for development of these skills, and tests that closely match th defined skills. If new materials and instructionally sensitive tests, such as the following were developed and used, then more students would probably achieve mastery, that is, meet the criterion level for passing the test:

    a. Teacher's Guide. Provide a guide for each course to assist teachers in their instructional efforts. The guide should explain each basic skill objective, state what content or experiences can be used to acquire the objective, suggest a set of instructional tactics for helping students acquire the content or processes subsumed by the objective, specify the intellectual capabilities and processes used in the objecitve and required by the test, give clear directions, and show the sample item format for the test items.

    b. Diagnostic test. This test would detect a learner's status in each component of a basic skill. The test would have subsections so that a teacher who wishes to know whether a learner has already mastered a particular part of a basic skill would then only administer the

tests related to this component. For example, a phonics test for sounds of initial consonants would be a part of one component for the basic skill of reading comprehension.

c. Mastery tests. These tests should be available to help the teacher monitor the student's progress towards mastery of each basic skill during instruction. These mastery tests would be directly related to the more general but limited focused skill definitions.

d. Student study guides. Each basic skill would be described in simplified language along with practice test items and their answers as well as a brief exposition of the skill and a set of selected textbook references for student use.

e. Self-instructional booklets. Materials that offer pupils opportunity to practice the basic skills. These booklets would be prepared after task analysis and would isolate the key prerequisites enroute to the basic skills.

11. Teaching and testing should be up-dated as new knowledge is acquired. Recent theory and research has shown that summarization rules can enhance retention of printed materials. When these rules were taught to elementary students, they also improved students' performance on tests of inference in the Califonria Assessment Program's Survey of Basic Skills.[29] This study which was conducted for this report, demonstrates the need for new instructional procedures and new diagnostic tasks.

This section completes our critical and contructed analysis of tests for use at the classroom. As a further basis for recommendations on policies for tests and testing at the state, school district, and classroom levels, we conducted a series of studies which are contained in Volume I of our two-volume Final Report at the California Policy Commission. Our recommendations are in our Executive Summary, which appears at the beginning of Volulme I and this volume. We hope that our report and its recommendations will help improve tests and testing in California, and if they do, they will accomplish our major objective, the amelioration of education for all students in California's schools.

References

1. Baker, W. and A. Cox (Co-Chair). Report to the Policy Committee on the University of California's Activities to Assist Underprepared Students. Part I: Descriptions of Trends and 1979-80 Activities; Part II: A Compendium of Ideas for Further University Activity. Berkeley, California: University of California, March 1981.
2. Balow, I. H. But, does CAP help school districts? California School Boards, September 1978, 37(7), 18-21, 34.
3. California Assessment Program. Technical Report. Sacramento, California: California State Department of Education, 1977.
4. California Assessment Program. Student Achievement in California Schools: 1979-80 Annual Report. Sacramento, California: California State Department of Education, 1980.
5. California Assessment Program. Survey of Basic Skills. Separate reports on Rationale and Content for Grades 1, 3, 6, and 12. Full Report: Part I, Survey Results; II, Program Diagnostic Displays; III, Interpreting and Using Test Results; IV, Skill Area Descriptions. Also separate Examiner's Manual for each test level. Sacramento, California: State Department of Education, 1980, 1982.
6. Glazer, R. The future of educational research. The Educational Researcher, 1982, 11(8), 16-17.
7. Kinzer, C. Field educators perceptions, attitudes, and values on testing reading: An international perspective. Paper prepared for the National Reading Conference's Annual Meeting, St. Petersburg, Florida, December 1982.

8. Kinzer, C. and R. B. Ruddell. Evaluations and testing of reading: Field educators' perceptions, attitudes, and values. Paper presented at the National Reading Conference's Annual Meeting, Dallas, Texas, December 1981.

9. Law, Alexander I. CAP - Statewide testing comes of age. California School Boards, September 1978, 37(7), 6-22.

10. McClung, M. S. Legal implications of intelligence and competency testing. Center for Law and Education, Cambridge, Mass.: Harvard University, 1979.

11. McNeil, J. and L. Donant. Summarization strategy for improving reading comprehension. In J. Niles and L. Miller (Editors), New Inquiries in Reading Research and Instruction. (Thirty First Yearbook of the National Reading Conference.) Rochester, New York: The National Reading Conference, 1982.

12. McNeil, J., H. Singer, and L. Furse. Curricular scope as a causal factor in reading achievement. Paper prepared for the National Reading Conference's Annual Meeting, St. Petersburg, Florida, December 1982.

13. Proficiency Assessment in California: A Status Report. Sacramento, CA.: State Department of Education, 1980.

14. Ruddell, R. B. Knowledge and attitudes toward testing: Field educators and legislators. Paper prepared for the National Reading Conference's Annual Meeting, St. Petersburg, Florida, December 1982.

15. Ruddell, R. B. and C. K. Kinzer. Test preferences and competencies of field educators. In J. Niles and L. Miller (Editors), New Inquiries in Reading Research and Instruction. (Thirty First Yearbook of the National Reading Conference.) Rochester, New York: The National Reading Conference, 1982.

16. Singer, H. The great educational earthquake of 1981: Reallocation of resources in California high schools to meet minimal competency requirements. Was the reallocation necessary? The California Reader, March/April 1982, 15(3), 13-19.

17. Singer, H. and M. J. Dreher. Preferences for testing and reporting test results in basic skills: Acquisition vs. learning from text, content-bound and content-free. In J. Niles and L. Miller (Editors), New Inquiries in Reading Research and Instruction. (Thirty First Yearbook of the National Reading Conference.) Rochester, New York: The National Reading Conference, 1982.

18. Singer, H. and M. J. Dreher. Attitudes towards testing and test results in achievement of basic skills. Final report, submitted to California Policy Seminar Project: Achievement Testing in Basic Skills. Berkeley, California: Institute of Governmental Services, 1982.

19. Singer, H., R. B. Ruddell, J. McNeil, and M. C. Wittrock. Testing Achievement in Basic Skills: Volume II: Description, Analysis, and Recommendations. (Final Report to California Policy Seminar.) Berkeley, California: Institute of Governmental Services, 1982.

20. Tyler, R. W., & S. H. White. Testing, Teaching and Learning. (Report of a Conference on Research on Testing, August 17-26, 1979) Washington, D. C.: National Institute of Education, 1979.

21. Utilization of the California Assessment Program by Local School Districts. Los Angeles, California: Southern California School Districts Superintendents' Council, November 1978.

22. Wittrock, M., H. Schultze. School and classroom evaluation. Final Report on California Seminar Project, Testing Achievement in Basic Skills. Berkeley, California: Institute of Governmental Services, 1982.

23. Gates, A. I. and G. L. Bond. Reading readiness: A study of factors determining success and failure in beginning reading. Teachers College Records, 1936, 37, 679-85.

24. National Assessment of Educational Progress. Reading in America: A perspective on two assessments. Reading Report No. 06-R-01. Denver, Colorado: National Assessment of Educational Progress, October, 1976.

25. Wirtz, W. Testing - The thermometer business. California School Boards, 1978, 37(7), 24-25.

26. Kirst, M. Speech to Technical Advisory Committee, State Department of Education, Sacramento, California, February 16, 1979 (multilith).

27. Coleman, J. The concept of equality of educational opportunity. Harvard Educational Review, 1968, 38, 7-22.

28. Singer, H. and I. H. Balow. Overcoming educational disadvantaged ness. In John Guthrie (Editor), Comprehension in Teaching: Research Reviews. Newark, Delaware: International Reading Association, 1981.

29. Hendrick, I. G. The early history of California state-printed textbooks. Southern California Quarterly, September 1964, 46, 223-238.

30. Dr. Steve Isaacs, Director of Research and Evaluation, San Diego Unified School District, supplied much of the information on school district evaluation.

31. Singer, H. IQ is and is not related to reading. In S. Wanat (Editor), Issues in Evaluating Reading. Arlington, VA.: Center for Applied Linguistics, 1977.

32. Bloom, B. S. Mastery learning and its implications for curriculum development. In E. W. Eisner (Editor), Confronting Curriculum Reform. Boston, Mass.: Little, Brown, 1971.

33. Metropolitan Reading Instructional Test, Elementary Form J. New York: Psychological Corporation. 1978.

34. Cook, W. W. The functions of measurement in the facilitational of learning. In E. F. Lindquist (Editor), Washington, D. C.: American Council on Education 1951.

35. Singer, H. and D. Donlan. Reading and Learning from Text. Boston, Mass.: Little, Brown, 1980.

36. Education Commission of the States, Department of Research and Information. States Activity - Minimum Competency Testing, 1980.

37. Caswell, H. and A. W. Foshay. Education in the Elementary School. Second Edition. New York: American Book, 1950.

38. The International Reading Association Board Position: Minimum Competencies. Newark, Delaware: International Reading Association, April 1979.

39. Coleman, J. S. Public vs. Private Education. Invitational paper read at the American Educational Research Association's Annual Meeting, Los Angeles, California, April 1981.

40. Klein, S., G. Fenstermacher, and M. Alkin. The Center's changing evaluation model. UCLA Evaluation Comment. January 1971, 2(4), 9-12.

41. Edelman, J. The impact of the mandated testing program on classroom practices: Teacher perspectives. Unpublished doctoral dissertation, University of California, Los Angeles, 1980.

42. Popham, W. J. Criterion-referenced measurement. Englewood Cliffs, N.J.: Prentice-Hall, 1978.

Appendix

Norm-Referenced and Criterion-Referenced Tests

To collect information for comparing achievement in basic skills, the State of California and school districts employ standardized, norm-referenced tests. Although these tests have some instructional applications, teachers are more likely to use criterion-referenced tests to assess the effects of instruction on achievement of specific objectives. A brief description of each of these types of tests will clarify their purposes and uses.

Norm-referenced tests

We can best describe a standardized, norm-referenced test by stating the criteria for evaluating it and describe how test-makers try to achieve these criteria:

Validity. This criterion consists of a set of procedures to determine whether the test measures what is says it is measuring. To satisfy this criterion, test makers start their construction of a test by defining some educational objective to measure, such as reading comprehension, arithmetic ability, and writing skill. They may further subdivide these objectives. For example, they may define reading comprehension as ability to identify the main idea, recall or recognize directly stated information, make inferences and interpretations, draw conclusions, and critically evaluate data contained in printed passages. They may divide arithmetic into fundamental arithmetic operations, problem solving ability, and sometimes knowledge of measurement. Writing ability may have separate subtests on composition, for example, a brief essay, language usage, punctuation, grammar, and spelling.

A way to determine if these definitions and subdivisions fit the curriculum is through a survey of courses of study and curriculum guides of school districts, county schools, and state educational agencies. From this survey, test publishers can arrive at instructional objectives common to the school system in the United States.[33] The degree to which the test objectives fit the curriculum is known as curricular validity.

We will illustrate this procedure for a reading test. After validating the objectives for the reading test, test-makers write passages that represent the reading content for each grade level and construct questions to assess comprehension of these passages. Samples of students at each grade level read the passages and answer the questions. Only those passages and questions that are appropriate for the grade level at which they occur remain in the test. All items that all students either fail or pass are excluded from the test. Usually only those items that correlate from .10 to .90 with the total test score are included on the test. Thus through sampling the domains of educational objectives and including a wide varied of objectives in each test level, standardized tests tend to achieve a normal distribution of scores because ability to attain these objectives are normally distributed in the population. Test publishers whose tests follow this test construction procedure can claim that their tests have not only curricular, but also content, and grade level validity. The next question is whether the test is reliable.

Reliability. Reliable tests consistently give the same results under the same conditions. It is tantamount to putting two thermometers into the same pot of heated water. If they give the same temperature reading, they are reliable. Test makers can construct two equivalent tests for this purpose or they can effectively do so by correlating the scores on the odd versus the even items of a test. (Actually they often use a statistical procedure which accomplishes the same result in a more effective way.)[5] Standardized achievement tests usually have reliabilities between .80 to .90, where .00 is no reliability and 1.00 is perfect reliability. Since standardized tests are not perfectly reliable, they have a chance variation which tests report as a standard error of measurement.

Standard error of measurement (SEM). The SEM of a test is an indication of how much a student's score is likely to vary if the student repeats the test. For example, if a student gets a score of 35 on a test which has a standard error of measurement of 2.0 points, on retaking the test several times under the same conditions, the student's score two out of three times could go as low as 33 or as high as 37. One third of the time, it could go beyond these limits.

The next question is what does a score mean? Of course, a score represents degree of ability to answer questions on a test that represents a sample of content. We can also determine what a score means by comparing it with scores that similar students earn on the same test. Test publishers provide this information in the form of test norms.

Test norms. To construct test norms, test-makers administer the test under the same directions and conditions (a part of the test's standardization procedure) to representative samples of students; usually they select their samples to represent the entire United States. They should describe in the test manual, which accompanies the test, the selection procedure for their sample of students, including how many students were in each selection category and the geographical areas they represent so that the consumer can determine the composition of the normative sample. In short, it tells the consumer the characteristics of the sample with which a particular student's score is compared.

Then the test-makers publish the distribution of raw scores, the number of items correct for this sample of students, arrayed from low to high, in a table of norms. Next to each raw score is the cumulative percent of students in the norm group at a particular grade level who got that score or a lower score. This cumulative percent is an equivalent score known as a percentile rank. For example, a student with a raw score of 48 may have a percentile rank equivalent score of 74, which means that this student's score equalled or exceeded 74 percent of the students in the norm group at this particular grade level.

In Table 6 is a simplified norm table taken from the Metropolitan Reading Instructional Tests. On the left side is the

Table 6

Condensed Norms for Grade 5 (Fall)[a]

Raw Score[b]	Percentile Rank[c]	Stanine[d]	Grade Equivalent[e]
60	99	9	12.9+
59	98		
58	97		
57	96		
56	94		
55	92	8	11.0
54	89		
53	86		
52	84		
51	82		
50	80	7	8.2
49	77		
48	74		
45	68	6	6.6
41	60		
40	58		5.7
35	50	5	4.9
30	40		4.2
29	38		
27	36		
25	30	4	3.9
23	26		
22	24		
21	22		
20	20	3	3.0
17	12		
16	10		
15	8	2	2.4
14	6		
13	4		
12	2		
11	1	1	1.9
10	1		

[a]These norms come from the Metropolitan Reading Instructional Test, Teacher's Manual for Administering and Interpreting, Intermediate, Forms JF and KI. New York: Psychological Corporation, 1978. Reproduced with permission.

[b]In the manual, raw scores occur consecutively from one to 60.

[c]Although percentiles go up from 1 to 99, the manual does not have one increment in raw score correspond with one increment in percentile ranking.

[d]The same stanine applies to all the raw scores and percentile ranks that lie between the lines.

[e]Grade equivalents are shown here for only every fifth raw score. However, the manual gives grade equivalents for each raw score.

distribution of raw scores on reading comprehension. In the next column is the equivalent percentile ranks for each raw score. In the third column are the stanine scores. The last column shows the grade equivalent scores.

A stanine (standard nine) is a division of the distribution of scores into nine equally spaced units (1 is lowest and 9 is highest). A particular percent of students is in each unit. For example, a student with a stanine of 5 earned a score on the test that is the range of scores earned by the middle 20 percent of students at that grade level who took the test. An advantage of reporting stanine scores is that they indicate a student's score lies within a range of scores. This range tends to include the chance fluctuations of scores, especially for scores in the middle of a stanine's range.

Grade equivalent scores indicate the average level of achievement earned by classes of students at successive grade levels. For example, suppose a reading test was given to a fourth grade class in October. Since October is the second month of the school year, the class was actually at grade level 4.2. At this time the class attained an average score of 30 on the reading test. The next May, when the class was actually at grade 4.8, its average score on the same test was 34. Through an interpolation procedure, the test-makers then plot what the raw scores and corresponding grade equivalents would have been between grades 4.2 and 4.8 had the class taken the test each month. The same procedure for computing grade equivalents was used at each grade level. The end result is the construction of

a table of grade equivalents for each raw score on the test. Hence, a tester or test consumer has to exercise caution in interpreting grade equivalent scores because the interpolated scores assume that development in reading achievement between October and May in each grade is linear.

Table 6 shows the stanines for the raw scores, percentile ranks within each stanine, and grade equivalent scores. From this table of norms, a teacher can interpret a student's raw score (number of items correct) on a test.

Interpretation of Raw Scores. After a student has taken a standardized test, the teacher can interpret his or her raw score by locating the score in the test's table of norms and obtaining the percentile, stanine, and grade equivalent for the score. For example, if a student in the fifth grade took the Metropolitan Reading Instructional Test in the fall and got 40 items correct on the reading comprehension test, the teacher would refer to the table of norms for this grade level (shown in Table 6) and would find that the percentile rank for a raw score of 40 is the 58th percentile, the stanine equivalent is 5, and the grade equivalent is 5.7. These equivalent scores place the student in the average range for the norm group. If a student is actually in sixth grade, but earns only a score of 48 on the test, then the student is reading below grade level, and vice-versa for a student who is in the fourth grade but also earns a score of 40 on the test. Thus, the teacher can determine how a student's score compares with a representative sample of students throughout the United States and whether the student's level of achievement in

comparison with national norms is at, above, or below or average for his or her actual grade placement.

Standardized test publishers usually claim that their tests meet the technical standards for standardized tests. But, as in all consumer activities, the rule of caveat emptor applies. Test users should read the directions for giving the test to determine whether they are explicit and clear, the test itself to determine whether it looks like it measures what it purports to measure, and the test's technical manual for information about the test's content and curricular validity, reliability, standard error of measurement, and description and criteria for the normative sample. They should also read reviews of the test in such standard reference works as O. K. Buros' Mental Measurements Yearbook and the Journal of Educational and Psychological Measurement.

A different type of test is the criterion-referenced or objective-referenced test. We shall briefly describe this test, also by explaining how to construct such a test.

Criterion or Objective-Referenced Tests

Test-makers, teachers or professional test publishers, construct a criterion-referenced test by specifying an objective in a particular instructional domain. Then they construct test items for this objective and specify a criterion score for passing the test. For example, a teacher at the third grade level may formulate an instructional objective that students should know how to use 5 syllabication principles for dividing words into syllables. The

teacher would select a list of 10 words for each principle, randomly arrange the words on a test page, and set a criterion that students should correctly divide at least eight words for each principle. Ideally, most students in the class should get a low score on the pretest, which would indicate that they needed to learn these principles. Then, if the students had the ability and wanted to learn the syllabication principles, and if the instruction was effective, then the pretest now given as a posttest will reveal which students had attained the criterion and had learned syllabication principles. In short, the criterion-referenced test assesses students on specific content, principles, or processes that teachers can teach and students can learn. It shows how students perform on the test before and after instruction. The test can show that students are alike before and after instruction in that most of them did not reach criterion before instruction, but most of them did reach criterion after instruction. If teachers can provide experiences for students to achieve the test's objective, allow time for at least 95 percent of the students to achieve this goal, and furnish instruction of high quality (instruction that is sequential, properly paced, has clear directions etc.), then it is possible for almost all students to reach a test's criterion level.

However, a teacher cannot claim that the students know any more than what the test specifically assesses. That is, a criterion-referenced test, unlike a norm-referenced test, usually does not sample an entire domain, such as every fiftieth word of a dictionary.

Hence, criterion-referenced tests usually can only indicate that students have or have not met a criterion score, that is, they have succeeded in attaining a particular objective. They may also report the percent of students at a particular grade level who meet this criterion. Thus, they also can furnish some norm-referenced data. But criterion-referenced tests do not determine students' level of general ability. That is what standardized, norm-referenced tests strive to do.

However, it is possible for a standardized test to contain a sample of items for a set of criterion-referenced objectives that do cover an entire curriculum. The Metropolitan Reading Instructional Tests use this strategy. Moreover, the Metropolitan not only provides norms but also the percent of students who pass each test item at the grade level at which the test-makers placed the items. Thus, the Metropolitan, as do some other commercially-published standardized tests, includes features of norm-referenced and criterion-referenced tests. Technical, criterion-referenced tests may also contain norms in the form of percent of students passing items at each grade level. Thus, the two types of tests tend to overlap.[42]

Use of Norm Referenced and Criteron-Objective References Test Data

Norm-referenced test data are best for comparing individuals, classes, school districts, and states provided, of course, that normative data are available in the test manuals for each of these categories. Although it is possible to use norm-referenced test data for making educational decisions at the classroom level, these

decisions are more likely to involve student placement, selection of students for educational programs, or certification of educational attainment rather than specification of the content or purpose of particular lessons. However, some standardized tests which indicate test objectives, items related to these objectives, and materials for teaching students how to attain these objectives are more instructionally relevant than those tests which do not furnish this information.

Conversely, criterion-referenced tests assess students' accomplishment of particular educational objectives. Teachers select items for these tests that they can teach and students can learn. Their purpose is to select items that all students can learn and objectives they all can attain. In short, criterion-referenced tests are appropriate for assessing instruction and learning in the classroom.

Classroom teachers also can and do construct tests that differentiate among students and use the test results for grading and for placement of students. However, most often, classroom teachers' purposes are to determine whether students need to learn something and, after instructing the students, assess whether they have learned what they taught. For this purpose, classroom teachers are more often likely to use criterion-referenced tests because they have instructional validity, that is, they are tests which measure objectives that are actually the basis for classroom instruction.

Thus, the uses of norm-referenced and criterion-referenced tests can differ. Norm-referenced data are best for comparisons of

performance. Criterion-objective referenced tests are most suitable at the classroom level whenever teachers want information on what students have learned through classroom instruction.

However, it is usually inappropriate and invalid to evaluate teacher effectiveness by using data collected by standardized, norm-referenced test or even data gathered by criterion-referenced tests. Evaluation of teacher effectiveness is a complex issue in itself which lies beyond the scope of this report.

Addendum to

Testing Achievement in Basic Skills, Volume II.

Changes in Pupil Proficiency Laws, Failure Rates, and Estimated Costs

of Mandated Remedial Proficiency Classes

Changes in Pupil Proficiency Laws

State law (AB 3408, Fall of 1976; and AB 65, 1977) requires school districts to provide remedial classes for students in grades 7 to 12 who fail one or more proficiency tests but does not reimburse the districts for any of these courses taught during the academic year. But, starting in 1980, school districts were required to offer remedial proficiency classes in the summer time or at breaks in year-long schools (Education Code 37252a) and were reimbursed for such classes (Education Code 42238 [g]). Consequently, school districts that offer remedial proficiency courses only when they are reimbursed do not have to transfer any resources from the 70 percent who pass the proficiency tests to the 30 percent who do not. Of course, the state funds provided for summer school remedial proficiency classes diminish the total state money available for support of other educational programs. In this sense, a reallocation of resources has occured and is still occuring.

Failure Rates

Since June 1981, California students have had to pass locally developed proficiency requirements and complete the local course of study requirements to receive a high school diploma. A report from the State Department of Education stated that in 1979 "an estimated 7.8 percent of the 1981 seniors would not graduate" [i.e., would not receive diplomas] (p. 29) due to failure on one or more proficiency tests.[1]

---

[1] Proficiency Assessment in California. Sacramento, California: State Department of Education, 1980.

Since some 300,000 students are in each grade, then 7.8 percent of 300,000 would mean that 23,400 students would not receive diplomas.

In the 1980 report, 50 percent of the districts stated that 40 or more percent of the class of 1981 had failed one or more proficiency tests; however, the students still had a year to retake and pass the tests they had failed. The State Department of Education found that "As of December 1980, 12 percent of all high school seniors were in danger of not graduating solely because of their failure to pass the proficiency tests." However, the August 1981 study showed that only 9 percent of California's twelfth graders failed to receive high school diplomas and that of all high school seniors, only 1 percent did not graduate because of failure to pass the proficiency tests."[2] In other words, 12 percent of California's students went into their senior year with one or more proficiency tests to pass. That is approximately 36,000 students. But only 1 percent, about 3,000 students, still had not passed all their tests prior to June graduation. What the statistics do not reveal is how many students dropped out of school in the 12th grade because they had repeatedly failed one or more proficiency tests.

Those who failed the test were more likely to be Black (41%) or Hispanic (29%) than White (20%) or Asian (18%). Those who passed were more likely to be fluent English-speaking students (96%) than limited-English proficient (4%).

Estimated Costs of Remedial Proficiency Classes

In December 1981, 24 percent of in-coming seniors, 45 percent of eleventh graders, and 61 percent of ninth graders had failed one or more proficiency tests. For the ninth graders, 38 percent had failed the reading test, 52 percent the writing

---

[2]Statewide Summary of Student Performance on School District Proficiency Assessments. Sacramento, California: State Department of Education, 1982. (Quotes are taken from p. 3. See also p. 4 and 5 for data reported here.)

test, and 46 percent the math test.

There are approximately 300,000 students per grade in California. Therefore, if all incoming ninth graders who had failed a proficiency test and took a remedial course to prepare for trying to pass the test the next time it was given, then 144,000 students would have enrolled in a remedial reading class, 156,000 in a remedial writing class, and 138,00 in a remedial math class. These figures are for ninth graders alone. If we added in the eleventh graders, we would double the number enrolled, and if we put in the 12th graders, we would have about two and a half times as many students enrolled.

Just taking the ninth graders who could enroll in remedial classes and putting them into modules of 25 per class, we would have 4500 remedial reading classes, 6200 remedial writing classes, and 5500 remedial math classes. If we took an average of ten classes per teacher per academic year, that would mean 1,620 teachers would be needed to teach these classes. At an average salary of $18,000 per teacher, it would cost $29,160,000 just for the ninth grade alone.[3] If we increased that by a factor of 2.5, it would be $72,940,000 that the state would have to pay in additional funds to school districts if the school districts offered these courses only in summer and paid teachers at a prorated salary for average teachers. If the courses were taught during the academic year, it would mean a reallocation of school district resources, and this reallocation would entail a shift in resources from the 55 percent of the students who pass all their proficiency tests in ninth grade to those who fail one or more of them. Although the state does not reimburse school districts for remedial classes offered in the academic year, the money for these classes still comes from the legislature as general funds for education, and money used for remedial classes is therefore not available for other educational purposes.

---

[3]Some school districts do not prorate academic year salaries for summer session instruction; instead, they pay a lower hourly rate. However, the state reimburses school districts for summer session courses according to a complex formula. (Education Code 42238 (g) (1))

Since school districts are more likely to offer courses only when they are reimbursed for them, they are likely to provide remedial classes mostly or only in the summer time. The State's bill for such courses could be as high as $72,940,000. These funds do not include any of the costs for the testing program itself nor for school personnel time required for mandatory conferences with parents of students not making sufficient progress towards meeting the district's standards of proficiency in the basic skills of reading, writing, and arithmetic.